Unit 1 Looking for patterns

Contents

Introduction

Welcome to M140! This module is about using data, usually in the form of numbers, to describe aspects of society and the environment so that we can understand, interpret and, on occasion, change the world around us.

A dictionary definition of statistics is: 'a science concerned with the collection, classification and interpretation of quantitative data'.

Statistics is a broad discipline, touching virtually all aspects of social and scientific activity. For example, many items in the news relate in some way to statistical questions, often with direct relevance to our own lives:

- Are average house prices going up or down in my area?
- How should I interpret league tables for choosing a school?
- Have the traffic control measures implemented in my town reduced car accidents?
- How is my pension calculated and how much can I expect to get?
- How safe and effective are the medical treatments I have been prescribed?

These questions, or others like them, might concern you – and all have a statistical component.

The word 'data' is regarded as plural. One observation is a 'datum'.

More generally, and on a much larger scale, statistical data are used to monitor the health of populations, the levels of development and inequality, climate change and the performance of global markets. Figure 1 shows one such use of statistics. The purpose of this module is to introduce you to the use of data in a variety of different situations, to enable you to take a more informed view of statistical data and statistical reasoning, and to teach you some of the statistical techniques that are used to make sense of data.

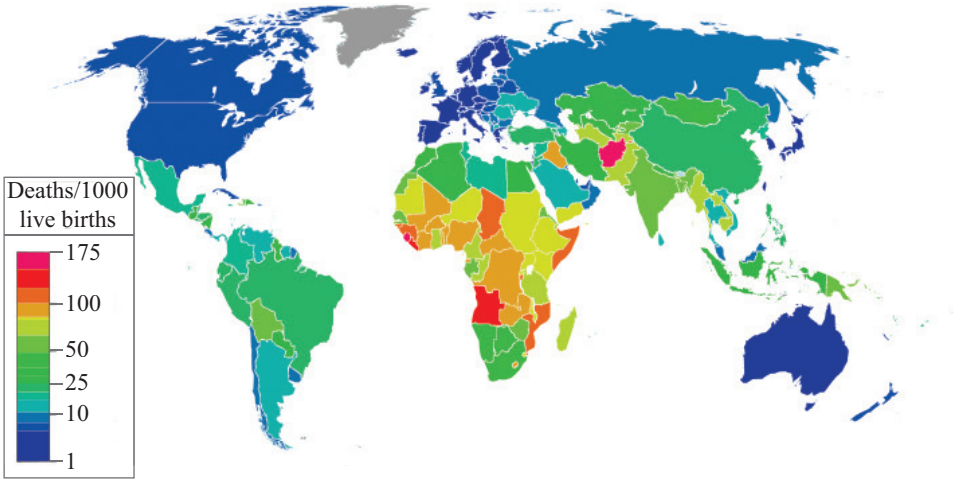


Figure 1 World infant mortality rates (2008)

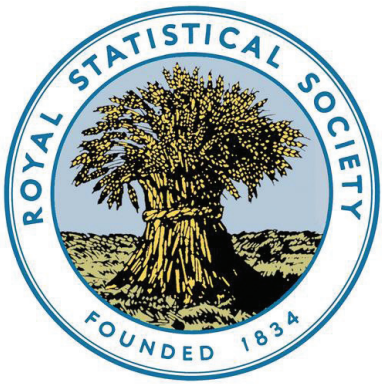


Figure 2 The wheatsheaf emblem

Threshing the data

The scope of statistics was the topic of some debate in the early days of the discipline. In 1834, the Statistical Society of London was founded, with a wheatsheaf as its emblem. This was inscribed with the Latin motto '*aliis exterendum*', translated as 'to be threshed by others', though the accuracy of this translation has since been disputed. This was interpreted to mean that while statisticians collect and present data (the wheat, in this analogy), they should leave the interpretation of data (the threshing) to others. This was hotly disputed at the time, the opposing view – generally accepted by

modern statisticians – being that statistics also involves interpreting and drawing inferences from data. Eventually, the Latin motto was dropped. The Royal Statistical Society, which succeeded the Statistical Society of London, still uses the wheatsheaf as its emblem, shown in Figure 2.

The teaching method we have adopted for the whole module is that statistics is best learned by *doing* rather than by *reading*. For this reason, this introduction has been kept short!

However, this being the introductory unit, we thought it useful to begin in Section 1 with a brief overview of the study materials and how to use them. In Section 2 we describe in general terms the process of statistical modelling. In Section 3 you will learn how to look critically at data, prepare it for further analysis, and undertake simple calculations. Sections 4 and 5 develop some simple tools for looking at data, and summarising some key aspects of a batch of data.

Sections 1 to 5 of the unit will require only pen, paper and a calculator, and you should have these at hand throughout. However, for handling larger datasets, and producing good graphs, a computer is essential. In Section 6 you will learn how to use the statistical package Minitab, including how to print/paste output for your assignments. So you will need access to your computer for that section.

Finally, Section 7 provides some advice and practice for the module assessments. Some of this will involve accessing online questions, so you will need access to the internet for this section.

1 Studying M140

This section contains some guidance on how to make the most effective use of the study materials and other resources. The M140 Guide also provides general guidance on studying the module, and you should read through this before you start Section 2.

1.1 The module components

Central to the module is the M140 website, where you will find copies of the M140 materials, the study planner, screencasts (audio–visual clips), practice quizzes, assessments and other information. It is worth checking the website regularly (say, once a week) for updates, news and corrections, and to drop in on the module forums to interact with other students.

The module software is provided separately on a CD.

Most of your studying will be based on the M140 units. Each unit of M140 is structured in a similar way. New ideas are illustrated by examples, and followed by activities for you to work through. Some of the new ideas are also demonstrated in the screencasts on the M140 website. The best way to learn statistics is to do it yourself, so it is important to work through the activities as you study the units. There are also exercises, usually given at the end of a section, which are there to give you extra practice should you feel you need it, and you can test your understanding of a unit by working through its associated practice quiz on the M140 website.

The Handbook contains some of the key ideas and formulas taught in M140, along with some useful statistical tables. You may find it helpful to keep the Handbook close by as you study the units, and add your own notes or examples

so that it becomes a convenient source of reference. Finding your way round the Handbook is easier than searching through the units, and using it to look up material from earlier units might come in useful at a later stage.

The Computer Book will guide you through the use of the module software, mainly through activities. The module software work will involve using the statistical package Minitab. The Computer Book also contains activities that use interactive computer resources on the M140 website, which are designed to reinforce ideas in the units.

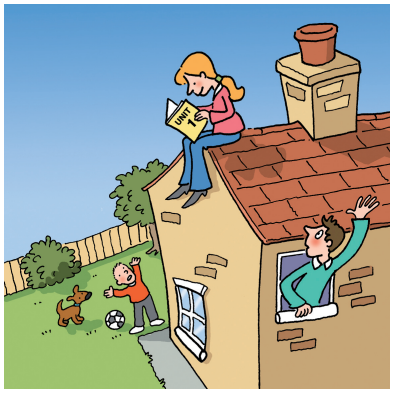
In the module material, there are icons to indicate where you will need to use the Computer Book, your calculator or an online resource. The corresponding icon for each of these is displayed here, from left to right. You might find these icons useful to plan your work, as you can see at a glance at which point these resources will be required.

The online work will include the screencasts and interactive computer resources, and you will be guided to the relevant ones as you work through the module material. The screencasts provide further demonstration of particular concepts and you may wish to review each of these straight after the relevant module material, or perhaps all together later on. The interactive computer resources are to be used with the relevant activities in the Computer Book.

Section 6 directs you to the parts in the Computer Book relevant to Unit 1. Note that you are also guided to part of the Computer Book at the end of Section 2, as you can choose to work through it at this point if you like. Section 7 of this unit will give you some information on working through the interactive computer-marked assignments (iCMAs), using the practice quizzes and answering questions in tutor-marked assignments (TMAs).

1.2 Studying effectively

The key to studying effectively is to plan your study time well ahead, and ensure you make suitable adjustments to fit your study into your life.



Jessica had to resort to extreme measures to find the time and space to study

Each unit is designed to take about 16 hours of study. This includes studying the unit (and any associated work in the Computer Book), reviewing the relevant screencasts, attempting the practice quiz and doing some of the end-of-section exercises. Some units may take you longer than others. You will need to allow some extra time for doing the assessment questions, and for other activities such as tutorials. It is important that you keep up with the study schedule in the study planner, which is on the M140 website, or you may find that you have run out of time to read the relevant units needed to complete an assignment by its cut-off date. So it's worth you trying to plan your study times, to fit them in with your other commitments, to make sure that you do not fall behind.

As you work through the units, it helps to have pen and paper ready to try out calculations for yourself, check you agree with the results in the unit, and to keep notes more generally. Remember, this module is all about *doing* statistics, not just *reading* about the subject.

It is important to remember that you are not alone with the study materials. Your tutor is there to help you with any statistical problems that you encounter. You should also raise with your tutor any matters to do with your progress, such as what you should do if you are worried that you may not complete an assignment or part of the module on time.

The module forums are another place where you can seek help, and also give help to others. If there's something you don't understand, then spend a few

minutes on it, make a note of it and perhaps return to it later. (You may find that it becomes clearer the second time round.) You could also see if one of the screencasts is relevant to whatever you are unsure about, as you may find it helpful to watch an example being worked through. But don't spend too much time puzzling over something without making progress. One step you can usefully take is to look on the module forums to see if any other students have met the same problem, and how they resolved it. You can also post your own query on a forum, or contact your tutor, and continue studying the rest of the unit in the meantime.

2 Patterns in data

The whole of M140 is about statistical modelling. In this section, the process is introduced in very general terms.

2.1 The statistical modelling diagram

In M140 you are going to spend a good deal of time looking at collections of numbers. Sometimes the data may display **patterns** that are incomplete or not immediately obvious. The objective will be to describe the pattern in each case and then, with the help of this description, to attempt to interpret this pattern in the context in which it occurs. The interpretation will often involve examining what the numbers suggest about a particular situation, and will sometimes enable us to make useful predictions about it. Sometimes you will find that what may have appeared to be a pattern was just a chance occurrence!

Here is a simple and familiar example to illustrate how patterns in data can be interpreted in context. More complex data and contexts will be introduced later.

Example 1 House numbers

Suppose that you are walking down a road and you observe that six consecutive houses on one side of the road are numbered 1, 3, 5, 7, 9, 11. You stop and ask yourself what number the next house will be.

Observations such as the house numbers of Example 1 are called **data** and the process of making **observations** is called **data collection**. Taken as a whole, the numbers 1, 3, 5, 7, 9, 11 are called a **batch of data** or a **dataset**.

In Example 1 you have done two things.

- You have **posed a precise question** to be answered: what is the number of the seventh house?
- You have **collected some data**, namely, the numbers of the first six houses: 1, 3, 5, 7, 9, 11. You hope to be able to use these observations to solve the problem.

The problem will be solved when a prediction has been made. To make this prediction, you need to complete a third task.

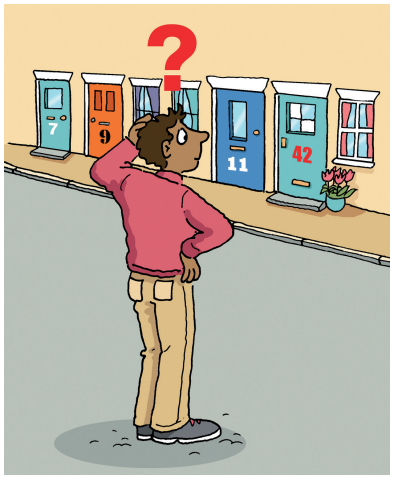
- You must **analyse the data** and see if you can find a pattern.

You have probably done this already and got the answer 13 because this number fits the obvious pattern. In doing this you may have used the rule 'the next house number is the next odd number in the sequence'. This rule describes the systematic pattern which, you believe, underlies the data: it is an example of a

model. The tasks 'pose a precise question', 'collect some data' and 'analyse the data' are important stages in the modelling process.

Activity 1 Predicting the next house number

Can you *guarantee* that the next house in Example 1 will be numbered 13? Explain your answer.



What you have done in the solution to Activity 1 is another important stage in the modelling process.

- You have **interpreted the model**.

In doing so, you did not just look at the pattern of the numbers. You also used your knowledge of what the numbers actually represented, that is, the fact that they were house numbers. This knowledge helps to interpret the pattern in the right 'real-world' context and in this case to suggest why the model might not apply to all the house numbers. It is essential to know the 'real-world' meaning of the numbers in a batch in order to interpret the data.

The four stages of the modelling process highlighted here are summarised in Figure 3.

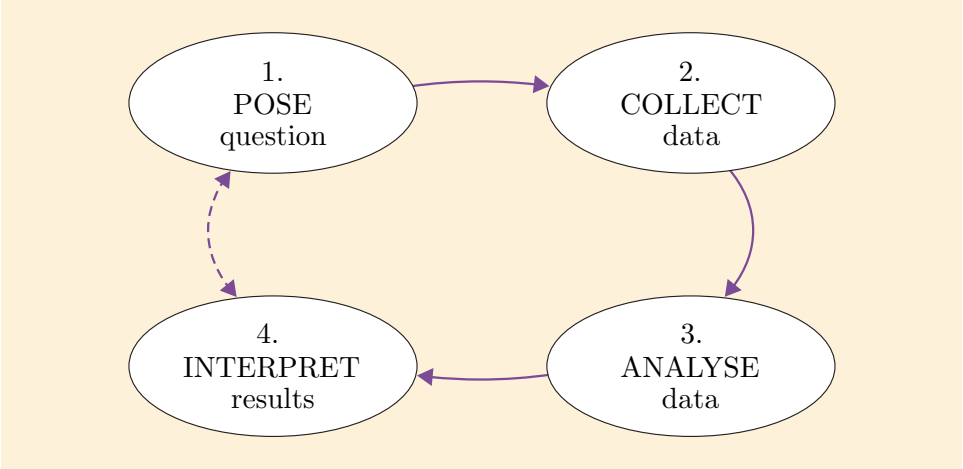


Figure 3 The modelling diagram

In fact, the house number example does not quite fit the modelling diagram since the question arose only after looking at the data – sometimes it needs some data to stimulate a question! The modelling diagram is really intended to cover the more systematic approach where the question posed comes first.

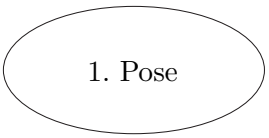
The **modelling diagram** will be referred to throughout this unit, and on many occasions throughout the module. This unit is concerned mainly with stage 3 – *analyse data* – particularly with the search for patterns in the data. However, the other stages will occasionally be mentioned.

It has already been said that an analysis should lead to a pattern from which predictions can be made. The pattern for the house numbers is that they are the odd whole numbers taken in ascending order. Normally, data will not exhibit such a clear regular pattern as that of the house numbers. The following example is more typical.

Example 2 Impact of fertiliser use on grain yield

Table 1 contains some data from an agricultural experiment to investigate the effect of fertiliser on the amount of wheat-grain produced. Let us presume that the question posed was:

How is grain yield affected by how much fertiliser is used?



and that the experiment was done to try and answer this question.

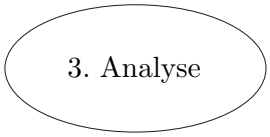
The 'pose' icon, based on Figure 3, is used to emphasise that this corresponds to the first stage of the modelling diagram. This and similar icons will appear at suitable points to emphasise which step of the modelling diagram is involved.

Table 1 Fertiliser use and grain yield

Fertiliser use (kg/ha)	0	25	50	75	100	125
Grain yield (tonnes/ha)	4.27	4.67	5.00	5.03	5.28	5.67

Here 'kg/ha' means that the fertiliser was measured in kilogrammes per hectare.

In this example we are interested in making predictions about how much grain might be produced by a given application of fertiliser. Specifically, we would like to know what would happen if *intermediate* quantities of fertiliser were used, such as 35 or 90 kg/ha, lying between the values for which the experiment was conducted.



Have a look at Figure 4, where we have plotted the data from Table 1. The quantity of fertiliser used (in kg/ha) is indicated horizontally and the grain yield (in tonnes/ha) is indicated vertically. Each experimental result is marked with a point, or dot.

This type of plot is called a **scatterplot**, as it helps to visualise a pattern (or lack of any pattern) in the scatter of data on the page.

In Unit 5 you will learn how to obtain such lines.

As you can see, the six points are roughly in a straight line. In Figure 5 we have drawn a straight line on the graph; this line represents a possible model for the data, which represents the observed pattern.

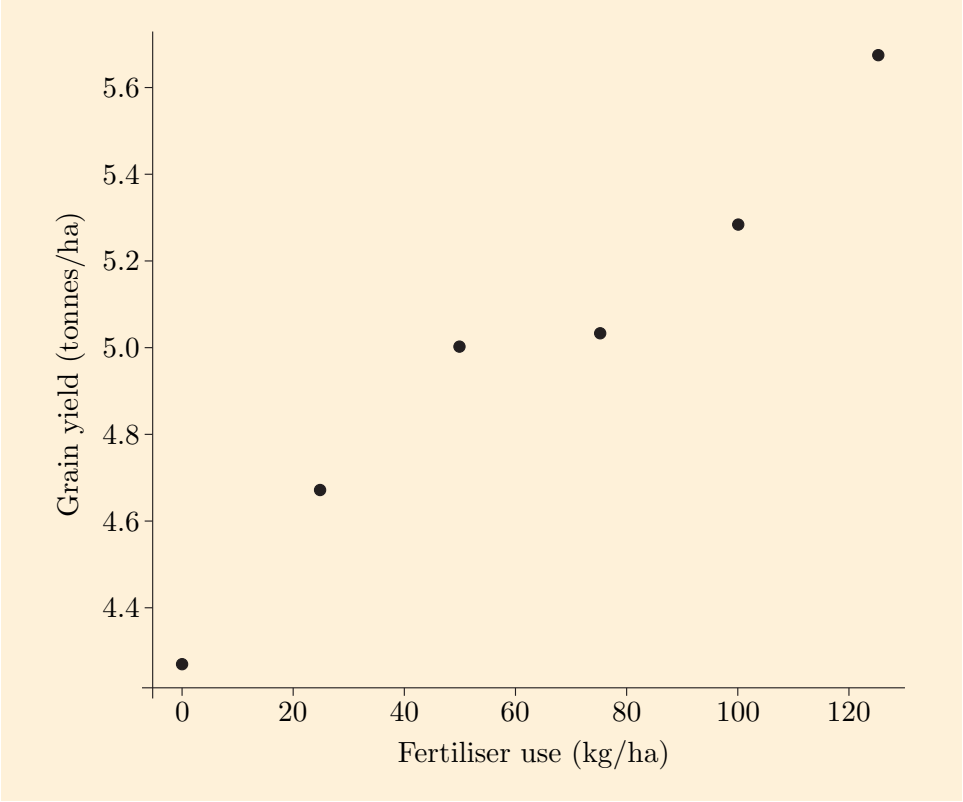


Figure 4 Grain yield by fertiliser use

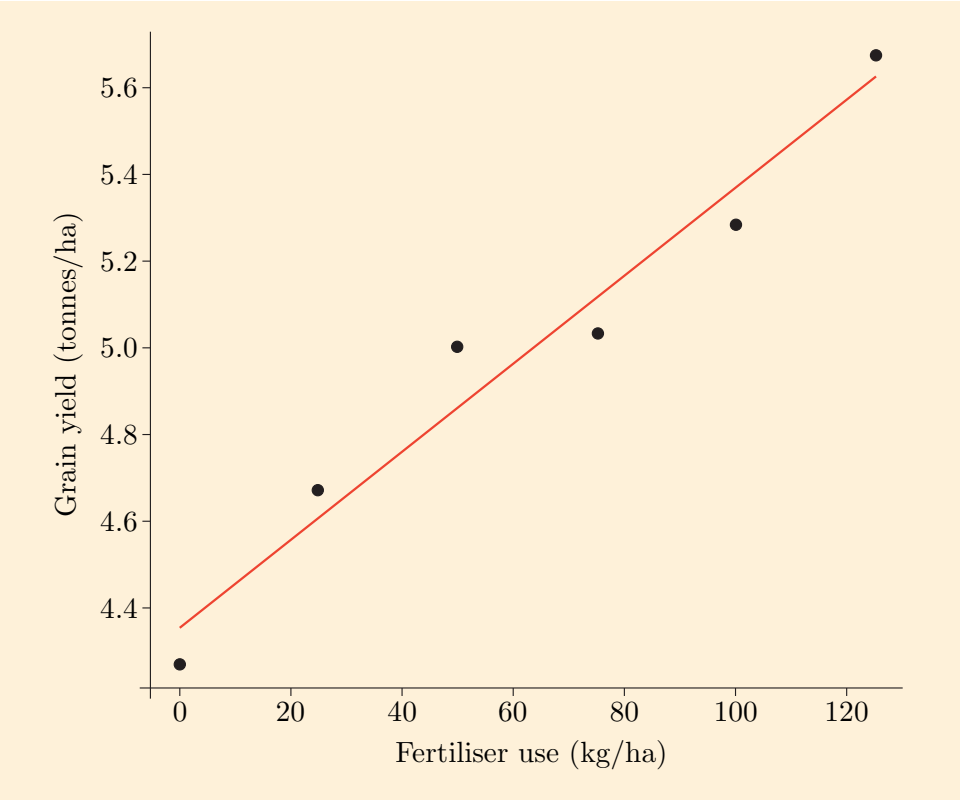


Figure 5 A possible model for the grain yield data

The straight-line graph represents a way of guessing a value of the grain yield from a value of the fertiliser use. This can be expressed by the simple formula:

$$\text{Grain yield in tonnes/ha} = 4.35 + 0.01 \times (\text{Fertiliser use in kg/ha}).$$

You will learn how to calculate this type of formula in Unit 5.

A value of the grain yield obtained from this formula can be regarded as a 'good guess' at what the value of the grain yield might have been for any particular quantity of fertiliser, if that amount of fertiliser had actually been used in the experiment. In this sense, it can be regarded as a **prediction** – actually a prediction about what might have happened at the time of the experiment, not about the future! For instance, for 35 kg/ha the formula predicts a yield of $4.35 + (0.01 \times 35)$, which equals 4.70 tonnes/ha, and for 90 kg/ha it predicts a yield of $4.35 + (0.01 \times 90)$, which equals 5.25 tonnes/ha. These should be good guesses at what might have been achieved in practice in these cases.

4. Interpret

However, for 50 kg/ha the formula predicts a yield of 4.85 tonnes/ha and for 100 kg/ha it predicts a yield of 5.35 tonnes/ha. These two predictions from the formula are clearly not exact because you can see from Table 1 that the actual yields were 5.00 and 5.28 tonnes/ha respectively. In general, predictions from the formula are not likely to be exact, but you can see from Figure 5 that the straight line is quite close to the data points and gives quite a good summary of the pattern in the data. However, it might not summarise the true situation so well for quantities of fertiliser lying outside the range of data here.

Example 2 is the subject of Screencast 1 for Unit 1 (see the M140 website).

2.2 Making use of patterns in data

In this subsection, the process of statistical modelling depicted in Figure 3 is illustrated by an example about *biodiversity*. Biodiversity is the variety of species of different living organisms that inhabit the Earth. This is often used as an

indicator of the health of an ecosystem, such as that represented in Figure 6. Therefore, it is of great interest to scientists to find out how many species of a particular plant or animal there are within a particular environment. The task is not easy, because in addition to the species we know about, there may be other species out there that we do not know about; new species are being discovered all the time.

Figure 6 A coral reef teeming with life

Example 3 How many large creatures are there in the sea?

This example is about the number of species of large creatures that live in the sea, where 'large' refers to a length of 2 metres or more for an adult. Thus the question of interest is:

How many large marine species are there?

One way to tackle this rather tricky question is to assemble data on the species discovered so far. The dataset used here lists the total numbers of large marine species so far discovered, in each year between 1829 and 1996. For example, in 1829 there were 101 large marine species known about; by 1996 the number was 217. (Source: Paxton, C.G.M. (1998) 'A cumulative species description curve for large open water marine animals', *Journal of the Marine Biological Association*, vol. 78, pp. 1389–1391.)

The next step is to plot these data, using a scatterplot, to see if a pattern is revealed. This plot is given in Figure 7.

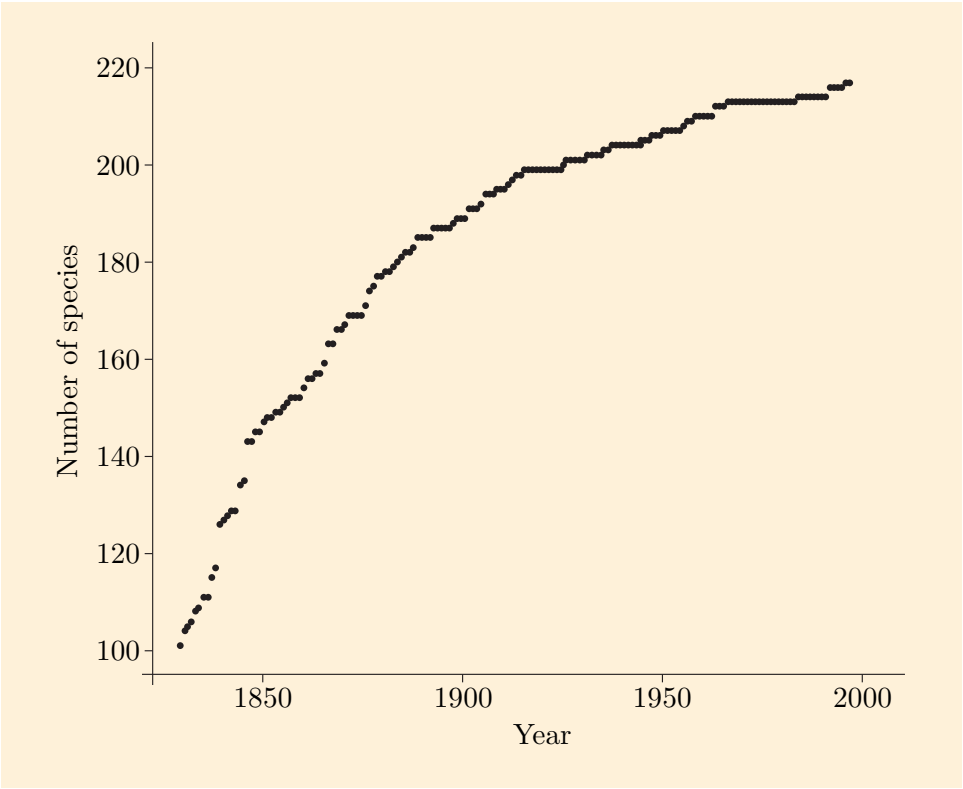
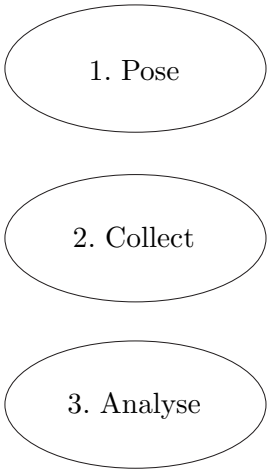


Figure 7 Number of large species discovered by year

The years from 1829 to 1996 are indicated horizontally, and the cumulative numbers of species known about by the end of each year (including those discovered that same year) are indicated vertically. This scatterplot is called the *discovery curve* for large marine species.



You will obtain this plot using Minitab in Section 6.

Figure 7 shows the numbers increasing each year, which is expected since a species known about one year will still be known about the following year. Also, while there are some wiggles in the curve formed by the points, it does appear to be quite smooth.

Activity 2 Interpreting the discovery curve: 1

Discuss briefly the following points relating to the pattern of the discovery curve in Figure 7.

(a) Is the discovery curve becoming more steep or less steep as time goes on?

(b) What reasons might there be for the change in steepness?

(c) What do you think might happen in the future (i.e. after 1996), based on the observations so far? What would you expect to happen?

In Example 2, a straight line provided an adequate representation of the fertiliser use and grain yield data. Here, a curved line is indicated for the discovery curve. One possible model for the data is shown in Figure 8.

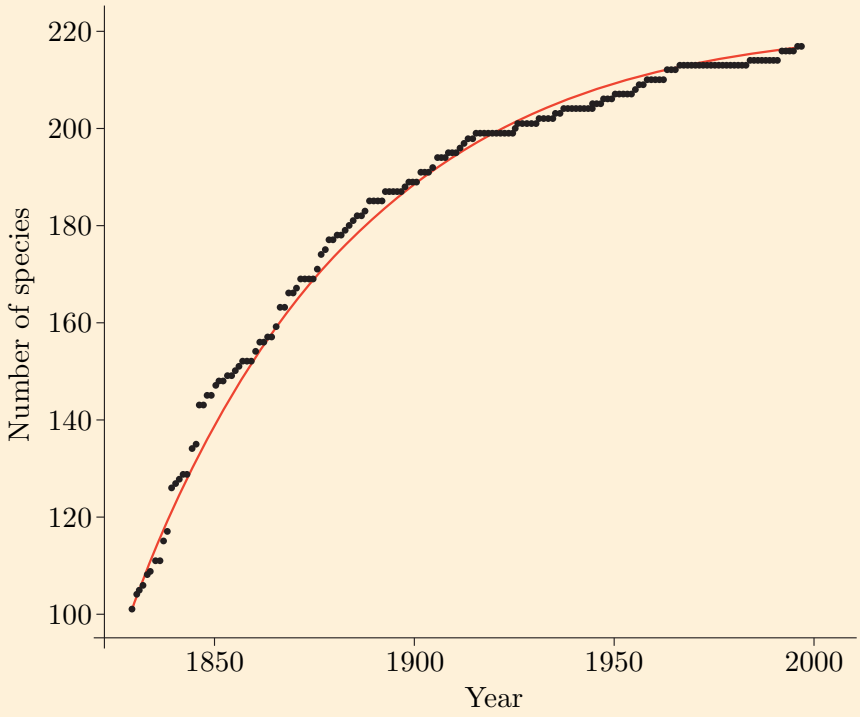


Figure 8 A possible model for the discovery curve

It is not the only such model for these data. How the model was chosen and fitted to the points is not important at this stage.

This model has been chosen so that the curve flattens out as time goes on, as would be expected with such data. The model goes through the data reasonably well, so it provides an adequate representation of the pattern in the data.

The final stage is to interpret the model in the context of the original question. This is the topic of the next activity.

4. Interpret

Activity 3 Interpreting the discovery curve: 2

The curve predicts the numbers of species known about each year, and flattens out in the future. In Figure 9, the curve has been drawn as before, but extended

until 2100, by which time it is nearly flat. The curve flattens out at a maximum value of 222.94, as indicated on the figure.

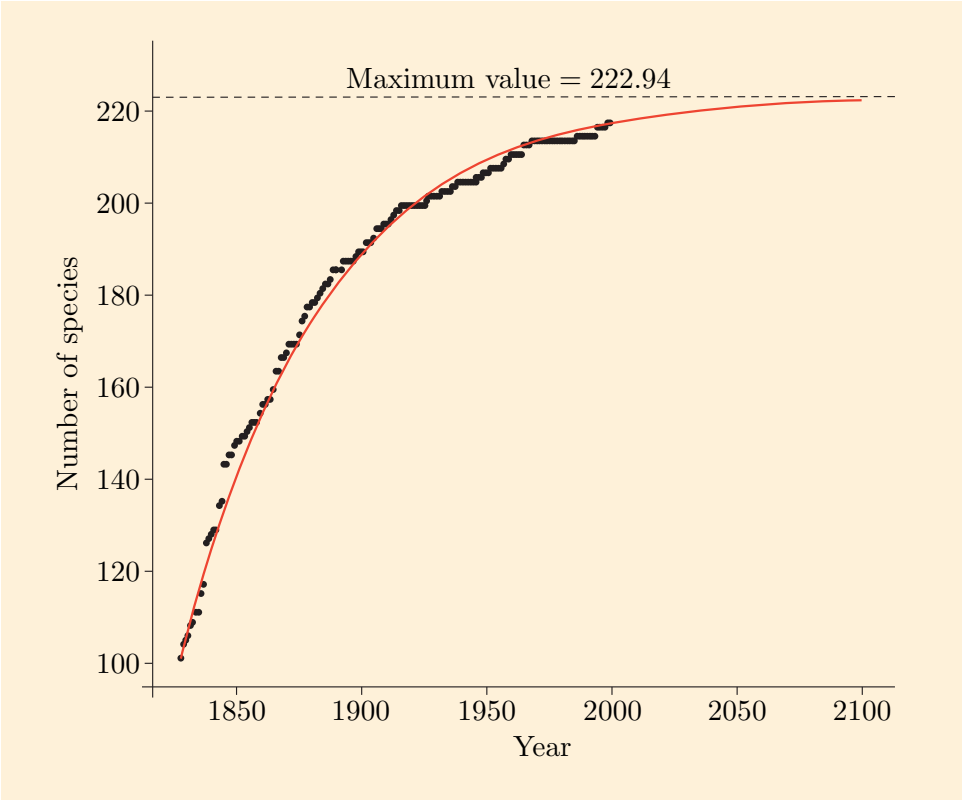


Figure 9 The discovery curve extended to 2100

(a) Use the information presented in Figure 9, together with your understanding of what the curve represents, to estimate the total number of large marine species. Bear in mind that this must be a whole number.

(b) In 1996 there were 217 large marine species known about. How many remained to be discovered?

(c) Briefly discuss what might happen if the curve through the points were changed slightly.

The yeti crab

New creatures are being discovered all the time. In 2005, the yeti crab (*Kiwa hirsuta*), shown in Figure 10, was discovered about 900 miles from Easter Island in the southeastern Pacific Ocean, by scientists working on the Census of Marine Life. It represents an entirely new family of marine organisms.

Figure 10 The yeti crab

In this section, you have learned about the process of statistical modelling, in general terms, through several contrasting examples. In particular, the modelling diagram will crop up throughout M140.

Exercises on Section 2

At the end of each section, you will usually find a few exercises to practise the skills you have learned. Unlike the activities within the sections, which you should work through as you come to them, you may find it helpful to leave working through these exercises until later – for example, when you are revising the unit.

Exercise 1 Your household budget

Statistical modelling is not limited to scientific investigations; it arises in everyday life as well.

Suppose that the question you have posed is:

How can I reduce my monthly household expenditure?

Write down what you might do to answer it, using the stages in the modelling diagram: Collect data, Analyse data, and Interpret results. That is, decide what data you would assemble, how you would analyse them, and how you would interpret your results to develop a plan for reducing your future expenditure.

Exercise 2 More interpretation of the discovery curve

Figure 8 shows the discovery curve, and a model for it, for large marine species. The curve becomes gradually less steep as time goes on.

(a) Interpret this pattern in terms of the average time between successive discoveries of large marine species, and how this average time varies with each new discovery.

(b) How, in general terms, might the interpretation from part (a) be used to predict the date by which the next large marine species might be discovered? How reliable would such predictions be?

You are now in a position to look at the Introduction to the Computer Book.

3 Preparing the data for analysis

There is an important preliminary part of data analysis which we have so far overlooked. It is not always possible to present data initially in such a neat and tidy fashion as we did in Section 2, particularly if the modelling diagram has not been followed and the data have not actually been gathered in a systematic way to answer a previously posed question. In practice, statisticians often have to use data that have been gathered with different (or no particular) aims in mind. A batch of data of this sort can contain errors, be incomplete in various ways and generally be 'not ready for immediate analysis'. It may even be slightly inappropriate for the questions asked of it. Some of these problems can occur even if the modelling diagram is being followed, but the problems are then usually 'accidental' in nature and should be less serious. Some of the problems can be corrected – others cannot. In this section, we consider some of the tools used for getting the data ready for analysis and some of the key issues relating to accuracy of the data.

3.1 Cleaning the data

The process of doing what you can to get the data ready for analysis is often referred to as **cleaning the data**.

Example 4 Petrol consumption and expenditure

For over 15 years, since acquiring his 1996 Honda Civic 1.4i, and with no particular reason in mind, a member of the module team noted down some information every time he filled the car's tank with petrol. He noted down the date, mileage, amount of petrol (in litres), cost per litre (in pence), and overall cost (in pounds sterling: *$*). At each petrol stop, the tank was completely filled up. Table 2 gives the data for the 2 years running up to the writing of this unit in early 2012; the dates in column two are in day.month.year format.

Table 2 Petrol consumption and expenditure data, Jan. 2010–Jan. 2012

Stop no.	Date	Mileage (miles)	Petrol used (litres)	Petrol price (pence per litre)	Expenditure ($)
1	18.01.10	112 350	46.01	109.9	50.57
2	18.02.10	112 616	44.98	110.9	49.89
3	04.03.10	112 954	43.49	111.9	48.67
4	24.03.10	113 269	45.13	115.9	52.—
5	20.04.10	113 564	41.76	119.9	50.07
6	—.05.10	113 857	44.25	119.9	53.06
7	—.06.10	114 123	40.70	116.9	47.58
8	—.06.10	114 469	43.38	119.9	52.01
9	05.07.10	114 823	41.60	115.9	48.21
10	11.08.10	115 091	41.17	112.9	46.48
11	27.09.10	115 360	45.83	112.9	51.75
12	09.10.10	115 706	43.14	114.9	49.57
13	02.11.10	115 984	44.26	115.9	51.30
14	25.11.10	116 202	37.95	116.9	44.37
15	06.12.10	116 542	45.30	119.9	54.31
16	24.12.10	116 795	40.88	121.9	49.84
17	10.01.11	117 071	37.49	126.9	47.58
18	22.01.11	117 417	44.89	126.9	56.97
19	02.02.11	117 736	40.59	126.9	51.52
20	10.02.11	118 030	33.02	126.9	41.90
21	17.02.11	118 368	42.87	129.9	55.69
22	25.02.11	118 748	47.47	126.9	60.25
23	04.04.11	119 064	36.—	129.9	47.70
24	11.04.11	119 419	41.57	130.9	54.42
25	20.04.11	119 773	45.53	132.9	60.51
26	01.05.11	120 134	43.75	134.9	59.02
27	11.05.11	120 481	37.51	133.9	50.23
28	20.05.11	120 794	40.37	13–.–	54.06
29	06.06.11	121 146	41.98	132.9	55.78
30	23.06.11	121 476	40.86	134.9	55.12
31	07.07.11	121 793	39.78	132.9	52.88
32	27.07.11	122 128	43.2–	132.9	57.43
33	08.10.11	122 436	38.87	131.9	51.27
34	03.11.11	122 786	40.47	128.7	52.08
35	19.01.12	123 108	44.97	130.9	58.87

(The numbers in the first column, 'Stop no.', in Table 2 are not really data but are used to number the rows of the table for easy reference. '—' indicates that digits are missing.)

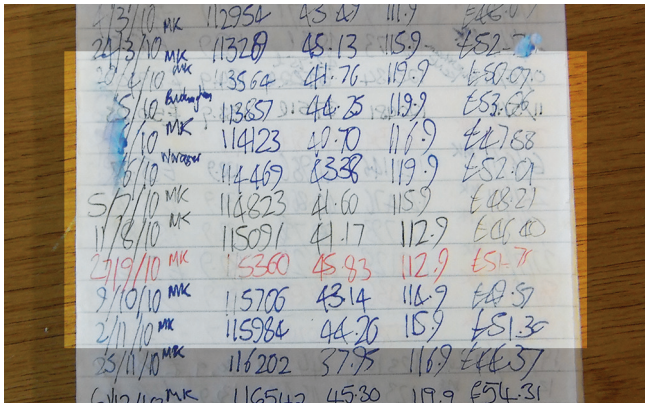


Figure 11 Notebook of data

The data in Table 2 present a few problems. In the 'Date' column the days in rows 6, 7 and 8, corresponding to stop numbers 6, 7 and 8, are missing: the notebook (see Figure 11) in which the records were kept got wet, and part of the entries written close to the edge of the page were illegible – there is nothing we can do about that, though we do know they were in May and June 2010.

The 'Mileage' column is complete. Some entries in the other three columns were partly illegible. This was the case for 'Expenditure' in row 4, 'Petrol price' in row 28, and 'Petrol used' in rows 23 and 32. Nevertheless, since the three quantities ('Petrol used', 'Petrol price' and 'Expenditure') are related, the missing entries can be recovered.

Activity 4 Calculating the expenditure

The relationship between 'Petrol used', 'Petrol price' and 'Expenditure' in Table 2 is given by:

$$\text{Expenditure (£)} = \frac{\text{Petrol used (litres)} \times \text{Petrol price (pence per litre)}}{100},$$

where the division by 100 is to convert pence into pounds. Given this relationship, use your calculator to obtain the expenditure corresponding to row 4, as accurately as your calculator permits.

The answer obtained for the converted price for row 4, namely *$*52.305 67, would look a little odd if inserted directly into Table 2 as it would have three more decimal places than the other values in the column! All this extra accuracy would be wasted when comparing these numbers with the others in the same column.

In addition, the extra accuracy is probably unjustified because the original data had only four digits in it.

Finally, the extra accuracy is unjustified in this context because the smallest unit of money is a penny, so it makes little sense here to quote all those decimal places. Thus, introducing all these extra decimal places would be to introduce **spurious accuracy** (or *spurious precision*), and that should be avoided.

Another type of spurious accuracy is to claim that a very precise statement is really meaningful, as in the following example.

It is common in mathematics to refer to this sort of accuracy as 'precision'. However, in this module, we use the two terms interchangeably.

Context is important here: exchange rates between currencies are often quoted to many decimal places, in order to convert large amounts accurately.

Example 5 Route planner

A popular online route planner gives the length of the fastest road route from Milton Keynes to Edinburgh as 342.20 miles, with an estimated travel time of

5 hours 31 minutes and 56 seconds.

Leaving aside the unusual (though scenic) route selected through the Scottish Borders (Figure 12), the overall distance is given to an accuracy of a hundredth of a mile, or about 16 metres, and the travel time to an accuracy of a second. Given the variation in distance and time according to which lanes you use, not to mention other factors such as traffic and weather conditions, this degree of accuracy is hardly realistic, and is not particularly helpful. In practice the most useful information you would probably take from this is that the distance is about 340 miles, with a travel time of about $5\frac{1}{2}$ hours.

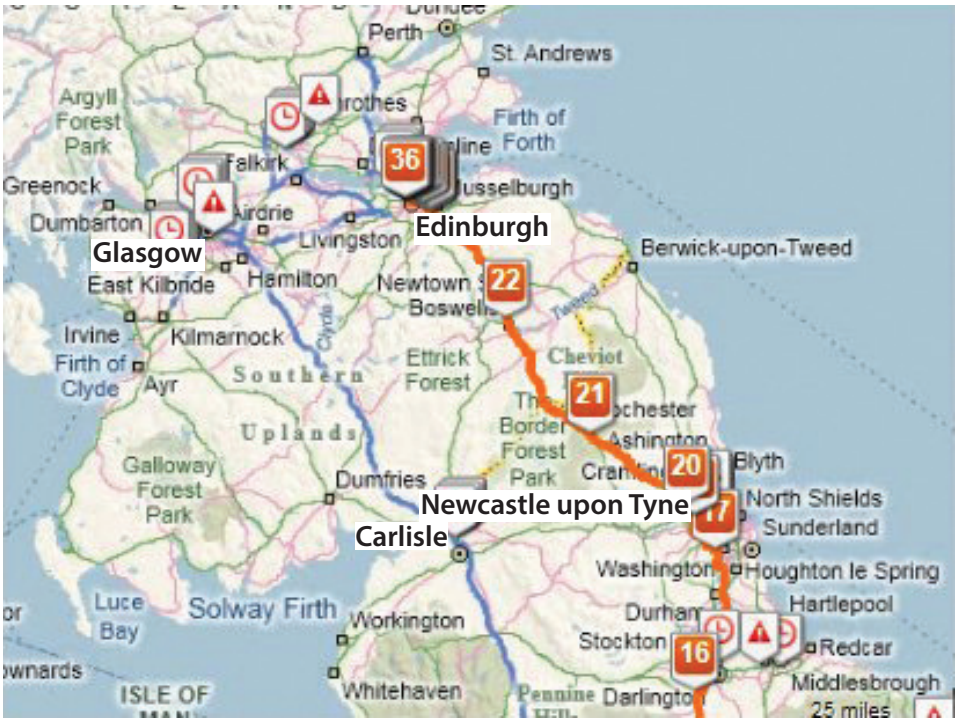


Figure 12 The last part of the recommended route from Milton Keynes to Edinburgh

As with the calculation described in Activity 4, the distance and time in Example 5 should be given to a suitable degree of accuracy to avoid giving a false impression. The next subsection discusses 'rounding' in order to achieve this.

3.2 Rounding

Rounding is used to give a value to a specified degree of accuracy. For example, if we wish to describe the petrol price (per litre) of 109.9 pence in Table 2 in terms of whole pence, the best approximation is clearly 110 pence (equivalently, *$*1.10) rather than 109 pence, 100 pence (*$*1.00) or any other value. Similarly, the price 110.9 pence in terms of whole pence is best described as 111 pence (or *$*1.11) rather than 110 pence or any other value. You can think of this approximation as being 'to the nearest penny'.

In M140 we shall use the following rule for rounding. (In other settings different rules are sometimes used.)

Rounding a number

To round a number, find the digit immediately to the right of where you want to round.

Round up if this digit is 5 or more, and round down otherwise.

Example 6 Round and round we go

Suppose we want to round 3.421 3604 to four decimal places. This is done as follows. We want to keep four decimal places, so we want to round at the position indicated by the vertical bar:

$$3.421\,3|604.$$

The digit to the right of where we want to round is 6. Since this is equal to or greater than 5, we round up to 3.4214. Similarly, 3.421 3604 rounded to six decimal places is 3.421 360.

The final zero is retained in 3.421 360 to show that we have rounded to six decimal places rather than to five decimal places.

Rounding before the decimal point is done in the same way. For example, if we want to round 176 354.67 to the nearest thousand, the position we wish to round to is indicated by the vertical bar:

$$176\,|\,354.67,$$

and the digit to the right of where we want to round is 3. Since this is less than 5, we round down to 176 000.

Example 6 is the subject of Screencast 2 for Unit 1 (see the M140 website).

The following activity involves rounding to avoid spurious accuracy.

Activity 5 The population of the world

In February 2012, surfing the internet for 'World population on 1st January 2012' produced one website that claimed:

> *As of January 1st, 2012, the population of the world was approximately 6 985 119 415 (6.99 billion).*

Clearly, a count of everybody on the planet was not undertaken instantaneously as the clock struck midnight in every time zone, so we can be confident that the number 6 985 119 415 is spuriously accurate (as acknowledged by the 'approximately' qualifying it).

(a) Round this population count to the nearest million.

(b) To what accuracy is the rounded value '6.99 billion', or 6 990 000 000, given in the quote? Is this value rounded up or down?

Rounding up – as practised in sheep-dog trials

Having practised rounding, it is time to round the values obtained in Activity 4, as this is what led us to consider the issue of accuracy in the first place.

Activity 6 Rounding the fuel consumption data

(a) Round the value obtained in Activity 4 to fit the values in the rest of the 'Expenditure' column in Table 2.

(b) Use the relationship between 'Petrol used', 'Petrol price' and 'Expenditure' given in Activity 4 to obtain the value for 'Petrol price' in row 28, and round it to fit the values in that column.

(c) Similarly, obtain the values of ‘Petrol used’ in rows 23 and 32, and round them to fit the values in that column.

These rounded values can now be written into the appropriate places in Table 2.

We have now completed the missing values in columns ‘Petrol used’, ‘Petrol price’ and ‘Expenditure’ of Table 2 and, since that was all we could actually do, the table has been cleaned up as far as it can be. All the numbers in any particular column are presented to the same degree of accuracy, so no further rounding is required. We shall assume that the degree of accuracy – no more and no less – is justified in every case. The only problems we are left with are some missing dates that we cannot resolve.

Before moving on to performing new calculations from the data, there is one issue which deserves comment, and that is the impact of ‘rounding errors’.

Example 7 Rounding error

If you apply the equation from Activity 4 to, say, the first row of Table 2, and calculate ‘Expenditure’ from the values provided for ‘Petrol used’ and ‘Petrol price’, and round the result to two decimal places, you get

$$\frac{46.01 \times 109.9}{100} = 50.564\,99 \simeq 50.56.$$

The symbol $\simeq$ means ‘approximately equal to’. The symbol $\approx$ can also be used.

This differs in the last decimal place from the entry for ‘Expenditure’ in row 1 of the table, which is 50.57. This is probably not due to a mistake in the data (though of course it could be), but is most likely due to a **rounding error**. The values for the price of petrol and the amount of petrol purchased that have been entered in Table 2 are most likely rounded values, and this is the reason for the slight discrepancy. For example, if the actual amount of petrol purchased was not 46.01 litres, but 46.014 litres, then the calculation would have been

$$\frac{46.014 \times 109.9}{100} = 50.569\,386 \simeq 50.57.$$

This now matches the value of ‘Expenditure’ in the table. However, the value of ‘Petrol used’, rounded to two decimal places, would still be 46.01.

Slight discrepancies due to rounding, typically only affecting the last digit, are frequent in tables, and are nothing to worry about. Large discrepancies, on the other hand, might indicate more serious errors. In calculating quantities to put into a table, it is important to reduce rounding errors to a minimum.

Reducing rounding errors

To reduce rounding errors in the final result, the full accuracy available should be kept in intermediate calculations, and the result should be rounded at the end.

Of course, you might be interested in the result of an intermediate calculation in its own right, in which case you should use the rounded value for that purpose, but still use the full accuracy to complete the calculation. The following activity illustrates what might go wrong if you round too early in a calculation, rather than keep the full accuracy available and round the final output.

Activity 7 A big rounding error

Calculate 19.4×23.4 and round the answer to a whole number. Next, round 19.4 and 23.4 to whole numbers first, then multiply them. What do you observe? Which result is correct?

In Activity 7, the rounding error 'grew' in the second calculation. In practice, the numbers you start with often include some rounding error, therefore it is important to know how to handle this in calculations so that the final result is not grossly inaccurate. The issue of how to handle numerical accuracy in calculations will be discussed further in the next subsection.

False security: a costly rounding error

During the first Gulf War, an American Patriot missile system (Figure 13) in Saudi Arabia failed to intercept an incoming Iraqi missile. The missile destroyed an American Army barracks, killing 28 soldiers and leaving many wounded. The cause of the failure of the missile system was eventually tracked down to a rounding error in the system's internal clock, which grew as time went on.

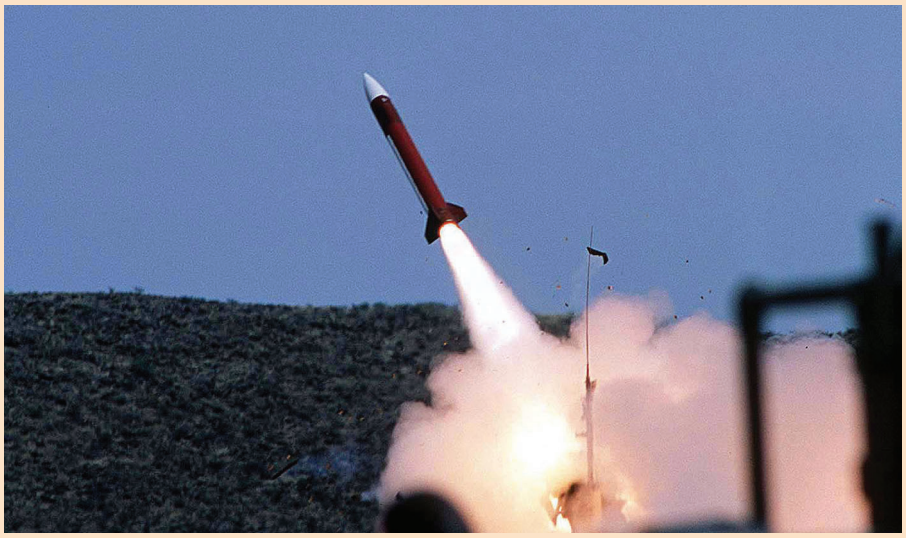
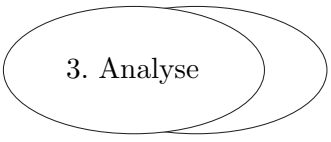


Figure 13 A Patriot missile

3.3 Calculating from the table

What can we actually calculate easily from the data in Table 2? One obvious question of interest is how expensive the car is to run.



This depends on fuel price and distance travelled, but also on the petrol consumption of the car. The next few activities and examples will take you through the steps required to calculate petrol consumption.

In the UK, petrol consumption is calculated from the following formula:

$$\text{Petrol consumption (miles per gallon)} = \frac{\text{Distance travelled (miles)}}{\text{Petrol used (gallons)}}.$$

(In other countries, fuel consumption is often calculated in metric units, frequently in litres per 100 kilometres, which corresponds more accurately to the meaning of 'consumption'.)

To begin with, note that from the mileage values in Table 2 we can certainly work out how many miles the car went between any two stops for petrol.

Activity 8 Distance travelled

How far did the car go in the month between 18.01.10 and 18.02.10? Use your calculator.

Note that we have departed from the modelling diagram in Figure 3, as the data were collected prior to posing the question. This can mean that the data you have are not ideal for answering the question you are interested in. Whenever possible, it's best to follow the modelling diagram.

Since the tank was completely filled up at each petrol stop, the petrol bought at each stop corresponds to the petrol used since the previous stop. Thus, you can calculate the petrol consumption between 18.01.10 and 18.02.10 using the data in Table 2. This is done in Example 8.

Example 8 Petrol consumption

From Table 2, there were 44.98 litres needed to fill up the tank on 18.02.10. This corresponds to the amount of petrol used since the previous fill on 18.01.10.

Next, we need to convert 44.98 litres into gallons. Now, 1 gallon equals 4.546 09 litres, hence

$$44.98 \text{ litres} = 44.98/4.546\,09 \text{ gallons} = 9.894\,216\,788 \text{ gallons}.$$

Remembering to keep maximum accuracy in intermediate calculations, we keep this ridiculously long number provided by a calculator for the time being in order to calculate the miles per gallon. From Activity 8, the car travelled 266 miles, so the fuel consumption in miles per gallon is

$$266/9.894\,216\,788 = 26.884\,391\,73.$$

Clearly, the eight decimal places give spurious accuracy to the result. But how should this number be rounded? In fact, there is a rule of thumb to guide this choice (discussed straight after this example) – and this rule indicates that the best choice here is to keep just one decimal place. Thus, our answer is that the car did an average of 26.9 miles per gallon between 18.01.10 and 18.02.10.

In Example 8, numbers obtained in the intermediate calculations contained a large number of decimal places. What degree of accuracy should be kept in the final result, to avoid quoting the final result with spurious accuracy? The answer depends on the numbers of **significant figures** involved.

The first significant figure of an unrounded number is its first non-zero digit, counting from the left. The next significant figure is the next digit (zero or other), and so on. Thus, for example, the third significant figure of the number 004 637 is 3; the fourth significant figure of the number 0.028 901 72 is 0. Rounding to a given number of significant figures is done just as before: for example, 004 637 rounded to two significant figures is 004 600 (or just 4600), 0.028 901 72 rounded to three significant figures is 0.0289.

Note that it's unlikely you'll ever encounter a number such as 004 637!

After a number has been rounded, it may not be possible to say how many figures are significant, unless you are told how it was rounded. For example, take the number 004 637 rounded to two significant figures, which is 4600. The two

significant figures are 4 and 6, and the last two zeros are not significant figures (that is, they are zeros because the number has been rounded). But 4600 is also what you would get by rounding 4601 to three significant figures, in which case the three significant figures are 4, 6 and 0. Or 4600 is what you would get by rounding 4600.1 to four significant figures, in which case the four significant figures are 4, 6, 0 and 0. Thus, a zero to the right of a rounded number may or may not be significant, according to how the rounding was done. However, when you see a measured quantity such as 4600 with no information about whether or how the number has been rounded, it is usually assumed that any zeros at the end are not significant.

Activity 9 Significant figures

(a) How many significant figures are there in 2460 if it has been rounded to the nearest ten?

(b) How many significant figures are there in 0.003 610 if it has been rounded to six decimal places?

(c) How many significant figures are there in 910 if you are not told whether, or how, the number has been rounded? (Give two options.)

(d) Round the numbers 208.3 and 0.098 3765 to three significant figures.

After this digression on significant figures, we return to the problem of how to round the result of an arithmetic calculation. The following rule of thumb provides a rough guide to this.

A useful rule of thumb

The 'output' result of a multiplication or division of several 'input' quantities, some of which have been rounded, should be rounded so that it has the same number of significant figures as the rounded input quantity with the *smallest* number of significant figures.

The rule is not infallible and sometimes the last digit can be wrong even if you know precisely how the numbers have been rounded, but it is a useful general guide!

Note that only the *rounded* inputs come into this. For example, to obtain a percentage, you need to multiply by 100: this 100 is not a rounded number, so does not count when working out how many significant figures to keep. If you do not know whether a number has been rounded, then assume it has.

Example 9 Rounding the petrol consumption

The numbers involved in the calculation for Example 8 were: the distance, 266 miles, which has three significant figures; the conversion factor from litres to gallons, 4.54609, which has six significant figures; and the amount of petrol used, 44.98 litres, which has four significant figures. It is reasonable to regard all these numbers as having been rounded. So the rounded number with the least significant figures is distance, with three.

Thus, according to the rule of thumb, you should keep three significant figures in the final result. In Example 8 this was 26.884 391 73, so keeping three significant figures gives 26.9 miles per gallon. (Ignore the decimal place when counting the number of significant figures.)

In the following activity, have a go at calculating the petrol consumption, in miles per gallon, between the next two petrol stops in Table 2.

Activity 10 Petrol consumption again

Calculate the petrol consumption in miles per gallon between 18.02.10 and 04.03.10, rounding your final answer to the correct number of significant figures. (Use the conversion factor of 1 gallon $= 4.546\,09$ litres.)

Whenever possible, it is better to do the whole calculation on your calculator without copying down intermediate numbers. This avoids making an error when copying, which is very easy to do. For example, in Activity 10, a suitable key sequence (provided your calculator has brackets) would be:

$$(112\,954 - 112\,616) \div (43.49 \div 4.546\,09) =$$

We can now proceed, by the method used in the solution to Activity 10, to calculate all the values of the petrol consumption between consecutive stops, using the data in Table 2. The results are presented in the last column of Table 3 – the answers have all been rounded to one decimal place. Also shown are the 'Period no.', where period 1 corresponds to the period between stops 1 and 2, period 2 to the period between stops 2 and 3, and so on, and the mileages at the petrol stops at the start ('Mileage reading 1') and end ('Mileage reading 2') of each period.

Table 3 Petrol consumption results 2010–2012

Period no.	Mileage reading 1	Mileage reading 2	Distance (miles)	Petrol used (gallons)	Expenditure ($)	Consumption (miles per gallon)
1	112 350	112 616	266	9.89	49.89	26.9
2	112 616	112 954	338	9.57	48.67	35.3
3	112 954	113 269	315	9.93	52.31	31.7
4	113 269	113 564	295	9.19	50.07	32.1
5	113 564	113 857	293	9.73	53.06	30.1
6	113 857	114 123	266	8.95	47.58	29.7
7	114 123	114 469	346	9.54	52.01	36.3
8	114 469	114 823	354	9.15	48.21	38.7
9	114 823	115 091	268	9.06	46.48	29.6
10	115 091	115 360	269	10.08	51.75	26.7
11	115 360	115 706	346	9.49	49.57	36.5
12	115 706	115 984	278	9.74	51.30	28.6
13	115 984	116 202	218	8.35	44.37	26.1
14	116 202	116 542	340	9.96	54.31	34.1
15	116 542	116 795	253	8.99	49.84	28.1
16	116 795	117 071	276	8.25	47.58	33.5
17	117 071	117 417	346	9.87	56.97	35.0
18	117 417	117 736	319	8.93	51.52	35.7
19	117 736	118 030	294	7.26	41.90	40.5
20	118 030	118 368	338	9.43	55.69	35.8
21	118 368	118 748	380	10.44	60.25	36.4
22	118 748	119 064	316	8.08	47.70	39.1
23	119 064	119 419	355	9.14	54.42	38.8
24	119 419	119 773	354	10.02	60.51	35.3
25	119 773	120 134	361	9.62	59.02	37.5
26	120 134	120 481	347	8.25	50.23	42.1
27	120 481	120 794	313	8.88	54.06	35.2
28	120 794	121 146	352	9.23	55.78	38.1
29	121 146	121 476	330	8.99	55.12	36.7
30	121 476	121 793	317	8.75	52.88	36.2
31	121 793	122 128	335	9.51	57.43	35.2
32	122 128	122 436	308	8.55	51.27	36.0
33	122 436	122 786	350	8.90	52.08	39.3
34	122 786	123 108	322	9.89	58.87	32.6

Note that the 'Petrol used' column in Table 3 has been expressed in gallons, and rounded to two decimal places. This column is included as an interesting intermediate result. However, the final petrol consumption figures were calculated using the unrounded values from Table 2.

In this section, we have seen that data always need to be carefully inspected and should be cleaned where necessary. Also, you have learned the technique of rounding. In the next section, we shall look at the values we have found for petrol consumption and try to display them in a way which may help us to see if there is any pattern in the data.

Exercises on Section 3

Exercise 3 Round and round again

Round 502.561 5297 to:

(a) six significant figures

(b) five decimal places

(c) the nearest whole number

(d) the nearest ten.

Exercise 4 The population of the world

In Activity 5, the population of the world was said to be 6 985 119 415 on 1 January 2012. Round this as follows, stating the number of significant figures in each case:

(a) the nearest ten thousand

(b) the nearest hundred thousand.

Exercise 5 Litres per hundred kilometres

In several countries, petrol consumption is calculated in litres per hundred kilometres. Using this measure and the data in Table 2, obtain the fuel consumption between stops 1 and 2. Take care to round the final result appropriately. Use the conversion factor of 1 mile $= 1.609\,344$ kilometres.

4 Pictures of data

We are now ready to look for patterns in the petrol consumption data introduced in Section 3. One way of finding these patterns is to represent the data pictorially in some way and then to look at the picture. In this section we introduce one method of picturing data: the stemplot. Before jumping in and discussing the stemplot, however, let us see if we can discern any pattern at all from the only type of display which we have used so far on our petrol consumption data, namely, the table.

Let us extract the petrol consumption data from Table 3, referring each value to the period of time to which the value applies. For ease of reference, the data we need are laid out in Table 4.

Table 4 Petrol consumption 2010–2012: 1996 Honda Civic 1.4i

Period no.	Petrol consumption (miles per gallon)	Period no.	Petrol consumption (miles per gallon)
1	26.9	18	35.7
2	35.3	19	40.5
3	31.7	20	35.8
4	32.1	21	36.4
5	30.1	22	39.1
6	29.7	23	38.8
7	36.3	24	35.3
8	38.7	25	37.5
9	29.6	26	42.1
10	26.7	27	35.2
11	36.5	28	38.1
12	28.6	29	36.7
13	26.1	30	36.2
14	34.1	31	35.2
15	28.1	32	36.0
16	33.5	33	39.3
17	35.0	34	32.6

Does inspection of the ‘Petrol consumption’ columns of Table 4 as they stand give a clear impression of any aspects of the petrol consumption figures themselves? Apart from showing that most values are in the ‘thirties’ and a few are in the ‘upper twenties’ or ‘early forties’, it is not clear that they do. Can we do

anything 'quick and easy' with the data which might reveal a pattern? Well, if we arrange the petrol consumption values in ascending order, for instance, then we obtain the order shown in Table 5, which should be read from left to right in successive rows.

Table 5 Petrol consumption values arranged in ascending order

26.1	26.7	26.9	28.1	28.6	29.6	29.7	30.1	31.7	32.1	32.6	33.5
34.1	35.0	35.2	35.2	35.3	35.3	35.7	35.8	36.0	36.2	36.3	36.4
36.5	36.7	37.5	38.1	38.7	38.8	39.1	39.3	40.5	42.1		

From Table 5, we can immediately see that there is no value below 26.1, seven values '35 point something', nothing above 42.1, etc.

What we would like is a way to display the sort of information that's shown in Table 5 directly in pictorial form. That is what the stemplot gives us.

4.1 Stemplots

The **stemplot** is a device for displaying numerical data in a pictorial structure. Our first example of a stemplot is shown in Figure 14, where we have plotted all the ordered petrol consumption data taken from Table 5.

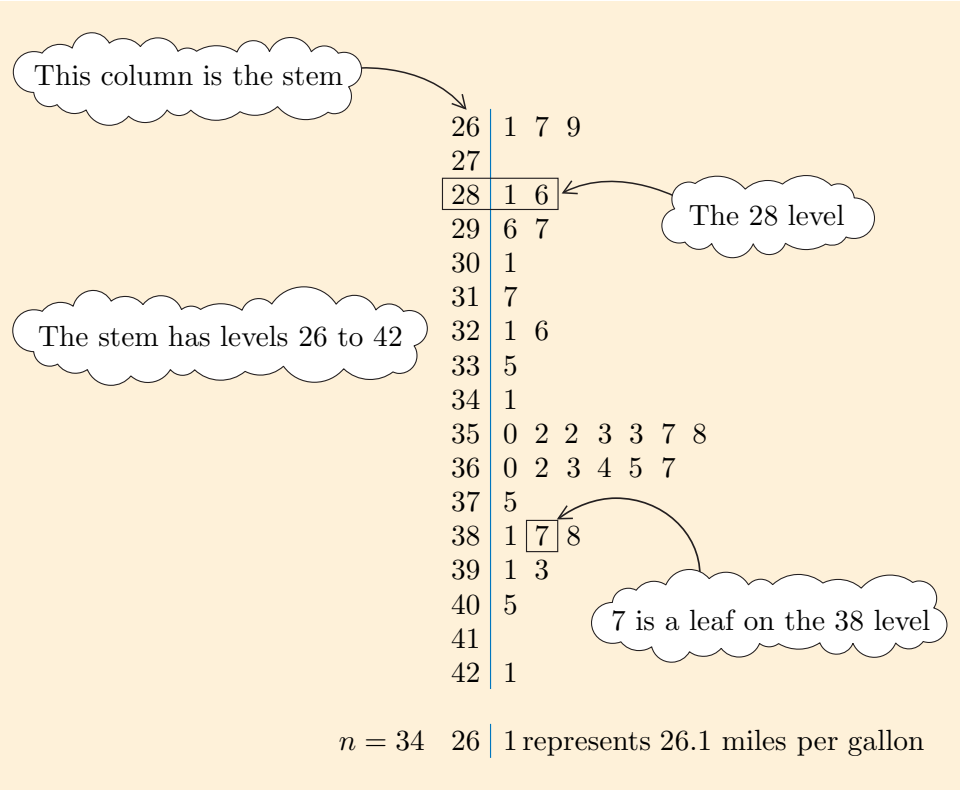


Figure 14 Stemplot of the petrol consumption data

Stemplots are also called *stem-and-leaf* plots.

The three basic elements in the stemplot are the **stem**, the **levels** and the **leaves**. The **stem of a stemplot** is the single column of figures lying to the left of the vertical line, arranged vertically downwards in increasing order. In Figure 14, this consists of all the whole numbers from 26 to 42 inclusive. We do not need any numbers less than 26 because the lowest data value (26.1) lies between 26 and 27. Similarly, we do not need any numbers greater than 42 because the highest data value (42.1) lies between 42 and 43.

Corresponding to each number on the stem is a **level** of the stemplot, which is named after that number. The level actually consists of everything occurring 'on

the same horizontal line' of the stemplot as the number used to name it. The '28 level' is surrounded by a box in Figure 14. Starting at the left-hand end of that box and moving to the right, we have first the number used to name the level (28), then a bit of the vertical line (which is used to keep the stem and leaves apart from one another) and finally two *separate* single digit numbers (1 and 6), each of which is referred to as a **leaf** of the stemplot. Each leaf on the stemplot corresponds to a single data value and the digit, from 0 to 9, used to represent it is derived from the data value itself. For example, the digit '7' on the '38 level', surrounded by a box in Figure 14, is a leaf on the '38 level' that represents the data value 38.7.

At the bottom of the stemplot are two pieces of information:

- The statement '$n = 34$', which indicates that there are 34 data values in the batch. Throughout this module we shall use the symbol n for the number of values in the batch, the **batch size**.
- '$26 \mid 1$ represents 26.1 miles per gallon', which is the rule (or key) enabling you to translate the combination of the level name (26) and leaf digit (1) into a data value as indicated above. This rule generally varies for different stemplots. It gives information about the units in which the data values are measured.

The next activity will give you some practice at reading a stemplot.

Japanese train timetables

Reading stemplots will come in handy if you go to Japan, where they are commonly used to represent train timetables, as in Figure 15.

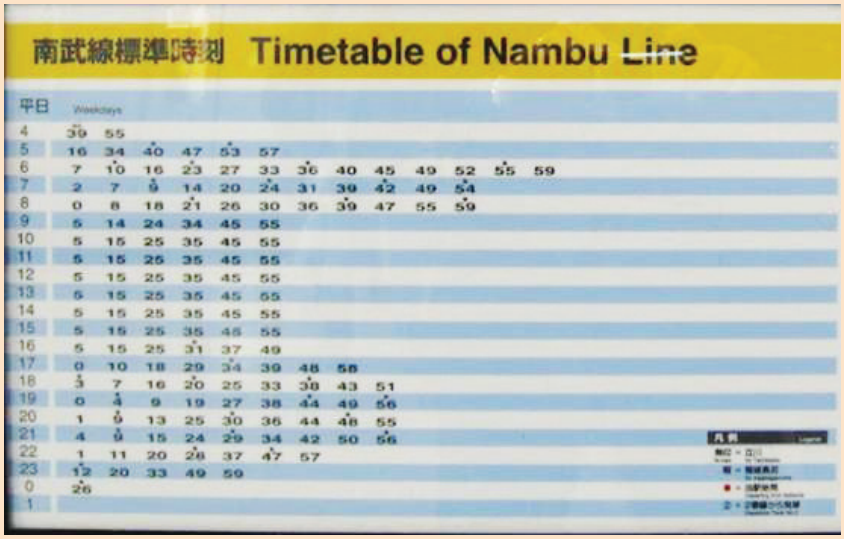


Figure 15 Train timetable for the Nambu Line, Japan

Activity 11 Reading a stemplot

Use the stemplot in Figure 14 to answer the following questions.

(a) How many levels does the stemplot have?

(b) Levels 27 and 41 have no leaves. What does this mean?

(c) How many levels have just a single leaf?

(d) How often did the value 35.3 occur in this batch? How is this represented on the stemplot?

One important thing to note about stemplots is that repeated values in the batch are represented by repeating the leaves. In Figure 14 the leaf 2 occurs twice at level 35, representing the value 35.2 occurring twice.

Another important feature is that, at each level, the leaf digits are *ordered*, increasing as you move away from the stem. This means that the data values are represented in order of *increasing* size as you move away from the stem at any particular level. As we shall see later, this is important when we want to determine certain things from a stemplot.

In Section 6 you will learn how to obtain stemplots using a computer.

In obtaining a stemplot by hand for a batch of data which has not been arranged in order of increasing size, as ours has been in Table 5, it is easiest initially to go through the data values one by one and draw up a stemplot by putting the leaves on each level in the order in which they actually occur. For example, if we had used the unordered values from Table 4 rather than the ordered values from Table 5, the initial *unordered stemplot* would have looked like Figure 16.

26	9 7 1
27	
28	6 1
29	7 6
30	1
31	7
32	1 6
33	5
34	1
35	3 0 7 8 3 2 2
36	3 5 4 7 2 0
37	5
38	7 8 1
39	1 3
40	5
41	
42	1
$n = 34$ 26	9 represents 26.9 miles per gallon

Figure 16 Unordered stemplot of the petrol consumption data

As a final step, the leaves on each level must be arranged in numerical order. For the petrol consumption data, Figure 14 would have resulted at this stage. The next activity will give you some practice at constructing a stemplot.

You have now covered the material related to Screencast 3 for Unit 1 (see the M140 website).

Activity 12 Counting moths

An ultraviolet moth-trap (as in Figure 17) was run at a location in North Buckinghamshire during many nights one September and the number of moths caught each night was counted. These nightly counts are given in Table 6 below.

Table 6 Nightly moth counts

47	21	16	39	24	34	21	34	49	20	37	65
67	21	37	46	29	41	47	24	22	19	54	71

Prepare a stemplot of these data, using the following steps.

(a) Explain why it suffices to use levels 1 to 7. Draw a vertical line, and write the numbers 1 to 7 as a column to the left of the line.

(b) Work along the first row of data and then along the second row, writing down each leaf at the appropriate level to the right of the vertical line. So, for example, for the value 47, write the leaf 7 at the 4th level. Once you have done this for all the values in the table, you will have constructed the unordered stemplot.

(c) Prepare a second stemplot with the leaves in correct numerical order.

(d) Label the stemplot in part (c) with the batch size and the rule to interpret each stem and leaf.

The following box summarises the key features of a stemplot.

The key features of a stemplot

- A stemplot breaks up the range of data values into a column (stem) of ordered levels corresponding to convenient equal intervals.
- The levels of a stemplot are labelled so that we know what each interval is.
- A stemplot represents data values in order within each interval (level) using single digit leaves.

Each leaf of a stemplot represents a data value whose first few significant figures are the level followed by the value for the leaf. For example, leaf b at level a represents a data value for which the significant figures start 'ab'.

So what does the stemplot actually give us which we did not have before? This is the topic of the next example.

Example 10 Interpreting the petrol consumption stemplot

The stemplot in Figure 14 provides a useful 'picture' of the data from which several features are apparent. For a start, the maximum and minimum values are easily read off from the plot (42.1 and 26.1 respectively).

Also, you could say that there is a 'pattern' and perhaps describe it as 'an obvious clustering around the 35 and 36 levels which diminishes gradually towards higher values and more steeply towards lower values, with perhaps a smaller clustering arising around the 26 to 29 levels'. This type of qualitative statement can often be made directly after simply *looking* at a stemplot. This would be much more difficult to do by staring at a table of data values, even an ordered one!

4. Interpret

This pattern can begin to suggest possible interpretations of the data. Typically, the fuel consumption lies around the mid to high thirties (miles per gallon), but occasionally there are periods of higher petrol consumption, with values around the mid to high twenties.

The interpretation of the stemplot in Example 10 was based on the following key factors, which help to describe what is usually called the 'distribution' of the data.

iscussed

Interpreting a stemplot

The *number* of leaves at each level tells us the number of data values in each interval. This indicates if there are any 'gaps' without many values, and where the maximum and minimum values are.

Comparing the numbers of leaves at the various levels indicates whether the data values 'cluster together' in particular regions.

The next activity will give you some practice at interpreting a stemplot.

Activity 13 Interpreting the moth-counts stemplot

In Activity 12, you prepared a stemplot for the data in Table 6. Comment on two aspects of the data revealed by your stemplot.

4.2 The median and the range

One reason for ordering the data on a stemplot is that it helps to find the **middle** of the batch. This is illustrated in the following example.

Example 11 The middle of the petrol consumption data

The petrol consumption data in Figure 14 show that the data values lie between levels 26 and 42, but cluster around the levels 35 and 36. Suppose now that we want to answer the question:

How many miles per gallon does the car do?

To answer this we should try and find a *single* number which is representative of the whole batch. The middle value of the data seems like it ought to be a good one to use (because it's in the middle!).

If we had an *odd* number n of data values, like 21 say, there would be only *one* middle value, which in that case would be the 11th value, taken from either end of the ordered data. There would be 10 values on each side.

However, in Figure 14 we see that $n = 34$, which is even. In this case there are *two* middle values, the 17th and 18th. Because there is no way to choose between them, we take the representative value to be half the sum (i.e. the average) of these two middle values. The quantity we define in this way is called the **median**.

The word *median* comes from the Latin word for middle.

If we look at Figure 14 and start counting leaves at the lowest value (that is, from the top of the stemplot), then the 17th data value is 35.3 and the 18th data value is also 35.3. So the median for the petrol consumption data is $(35.3 + 35.3)/2 = 35.3$, in miles per gallon. It so happens in this case that the median has a single decimal place; in other situations it might not, in which case the median would be rounded to match the rest of the data.

The median is an important quantity in statistics, which merits its own special box.

The median

The median of a batch of data is the middle of the ordered batch.

- If the batch size is odd,

  $$\text{median} = \text{middle data value}.$$

- If the batch size is even,

  $$\text{median} = \text{average of the two middle data values} = \frac{\text{sum of the two middle data values}}{2}.$$

The median should be rounded to the same level of accuracy as the original data.

We round the median to the same level of accuracy as the original data so as to avoid introducing any spurious accuracy. The median of a batch will often be used in this module as a single representative figure for a batch of data. It provides a measurement of the typical value or **location** of the batch.

Activity 14 The median moth count

Use the ordered stemplot you obtained in Activity 12 to find the median of the moth-trap data.

Note that if the batch size is even, the median can take a value that is not equal to *any* of the data values in the batch – this happened for the moth-trap data in Activity 14.

One of the big advantages of the median is that it is not affected by unusually large or small observations in a batch of data. Such extreme observations are by their very nature not typical, and so it is desirable that they should not affect our measure of what is a 'typical' value. This issue is illustrated in the next example.

Example 12 Average salaries for high-paid occupations

The Annual Survey of Hours and Earnings (ASHE) is undertaken annually by the Office of National Statistics and records the occupations and pay for a 1% sample of employees in the UK. The survey does not cover self-employed people and does not take into account non-salary rewards such as bonuses and share options. Thus it does not provide a full picture of income inequality in the UK (Figure 18). Table 7 shows the median annual salaries for the top 20 highest-paid categories of employees, derived from the 2011 ASHE data by a national newspaper. For example, the top category represents 'Heads of major organisations', and the value *$*114 550 represents the median salary among all employees surveyed who were in that category.

Table 7 Median salaries for the 20 highest-paid occupations in 2011

44 160	44 870	45 260	45 360	45 410	45 420	47 250	47 900	48 450	51 090
52 900	53 740	56 800	58 750	59 300	60 100	74 440	78 180	82 960	114 550

Figure 18 Income inequality provides a focus for political protest especially at times of economic crisis, as with the 2011 'Occupy' movement pictured here

Suppose now that we wish to represent these data using a stemplot with levels 4 to 11. To represent the data on a stemplot, we retain one leaf digit, and thus drop

the last three digits of each number. Numbers for which the last digits have been dropped without rounding are said to have been **truncated**. Thus, 44 160 and 44 870 are both truncated to 44. The stemplot based on the truncated data is shown in Figure 19. (It is customary, when drawing a stemplot, *not* to round the values. This makes it easier to vary how the stem is chosen.)

4	4 4 5 5 5 5 7 7 8
5	1 2 3 6 8 9
6	0
7	4 8
8	2
9	
10	
11	4
$n = 20$	4 | 4 represents £44 000

Figure 19 Stemplot of median salaries for highest-paid occupations

The single value '$11 \mid 4$', representing *$*114 000, is noticeably separated from the rest of the data. Such a value is called an **outlier**. (Precisely how far away from the main body of data an entry has to be before it can be deemed to be an outlier is somewhat arbitrary.)

Since there are 20 data points, the median is the average of the 10th and 11th largest values. From the stemplot, these are *$*51 000 and *$*52 000, so the median (calculated from the stemplot) is *$*51 500; rounding this to the same accuracy as the rest of the data on the stemplot gives *$*52 000.

The original values are *$*51 090 and *$*52 900, so the median calculated from the data is *$*51 995. Rounding this value to the same level of accuracy as the data (which appears to be given to the nearest *$*10) gives *$*52 000, the same as the median calculated from the stemplot. However, generally, the median calculated from the data and from the stemplot may not coincide exactly, due to the truncation of the data to construct the stemplot.

Note that the high value of the outlier does not influence the calculation of the median. You would obtain the same result if all heads of major organisations were paid millions (which, of course, some are). Similarly, the result is not influenced by the lowest value in the dataset. The median is said to be **resistant** to outliers.

To make a stemplot more compact, outliers are often listed separately, as in Figure 20 for the median salary data.

4	4 4 5 5 5 5 7 7 8
5	1 2 3 6 8 9
6	0
7	4 8
8	2
HI	114
$n = 20$	4 | 4 represents £44 000

Figure 20 Representing outliers on a stemplot

If there are several outliers, they are listed in order of size. Note that the notation used for them is 'level number followed by leaf digit'. For instance, in Figure 20, 114 represents *$*114 000. The label *HI* indicates that this outlier is a **high outlier**; that is, that its value is significantly *greater* than the values in the main body of the stemplot. High outliers are always listed *below* the stem, in increasing order.

Low outliers are values of the data that are significantly *less* than the values in the main body of the stemplot. The low outliers are always listed *above* the stem, in increasing order, and are labelled *LO*. (There are no low outliers in Figure 20.)

Activity 15 Low salaries

The ASHE (see Example 12) also collected information on part-time and low-paid jobs. Table 8 gives the median annual salary paid for the twenty lowest-paid occupations in 2011. (Many of these involve primarily part-time work.) The two lowest values correspond to school midday assistants and school crossing patrol attendants.

Table 8 Median salaries for the 20 lowest-paid occupations in 2011

2 190	3 630	5 530	5 660	6 060	6 500	6 620	6 680	6 980	7 030
7 870	8 330	8 360	8 480	8 740	8 840	9 120	9 600	9 900	9 980

(a) Construct a stemplot for the data in Table 8 with levels 2 to 9 (i.e. truncate the data after the second digit, without rounding). Treat the two lowest values as outliers and list them separately.

(b) Calculate the median for these data from the stemplot.

(c) Would the median change if the two lowest-paid occupations each had a rise of *$*1000? Explain your answer.

Another important property of a batch of data, which certainly strikes you straight away when you look at a stemplot, is its **spread**, or **scatter**. The stemplot gives a good *picture* of the spread but it would be convenient to have a single *number* which represents it, in the way the median represents the overall location of the values in a batch. Such numbers are called **measures of spread**. The simplest single-figure summary of the spread is the distance between the two extreme values, that is, the distance between the maximum and minimum. This is called the **range**.

The range

The range of a batch is the distance between the two extreme values. It can be calculated from the formula:

$$\text{range} = E_U - E_L,$$

where E_U is the **upper extreme** (the largest value, or maximum) and E_L is the **lower extreme** (the smallest value, or minimum).

Example 13 Calculating the range

The range of the batch of petrol consumption data in Figure 14 is calculated from the two extreme values, which are:

$$E_U = 42.1 \quad \text{and} \quad E_L = 26.1.$$

Therefore the range, in miles per gallon, is:

$$E_U - E_L = 42.1 - 26.1 = 16.0.$$

Note that there is no need to round the value, as subtracting (or adding) values does not increase the number of decimal places. However, also note that the result is reported as 16.0 rather than 16, to indicate that the result is accurate to the nearest 0.1 miles per gallon (and so has three significant figures).

Activity 16 The range of top salaries

Calculate the range of the batch of median salaries for the 20 highest-paid occupations in the stemplot of Figure 19.

As with the median, the range can be calculated directly from the original data, rather than from the stemplot. Because the range depends only upon the extreme values, it is very easy to calculate.

However, unlike the median, the range is not resistant to outliers. It is quite possible for a batch of data to have most of its data values clustered in quite a narrow region, but to have, say, a single high outlier and a single low outlier, both very far from all the other values. Then, as the two extreme values are far from the rest of the data, the distance between them gives little information about how tightly clustered all the rest of the data values are. The same situation often applies even if only one of the extreme values is an outlier. Units 2 and 3 introduce better measures of spread.

In this section, the concept of a stemplot has been introduced. You have seen how it can be used to calculate the median and we have looked briefly at how to interpret a stemplot to describe the distribution of the batch of data. In the next section, you will learn how to extract more information from a stemplot about one aspect of the distribution of a batch of data – its shape.

Exercises on Section 4

Exercise 6 Shot-put championship results

Table 9 gives the shot-put results for 15 senior male athletes in a UK athletics championship, in June 2011. The athletes were grouped into two pools, A and B. The results are in metres.

Table 9 Shot-put championship results (metres)

Pool A	12.57	12.75	15.43	16.27	16.40	16.70	18.05	
Pool B	11.98	12.37	13.87	13.91	14.38	15.10	15.28	16.47

(Data source: UK Athletics)

(a) Construct a stemplot for all 15 values with levels 11 to 18.

(b) Calculate the median from the stemplot.

(c) Obtain the extreme values and the range from the stemplot.

(d) Briefly comment on the distribution of this batch of data. Are there any potential outliers?

Exercise 7 Wooden toy prices

Table 10 shows the prices, in increasing order, of 28 different wooden toys costing under *$*20 listed by an online retailer in February 2012.

Table 10 Wooden toy prices (*$*)

1.21	1.21	4.05	5.99	6.04	7.25	7.50
7.99	7.99	8.49	8.75	9.10	9.40	9.43
9.50	9.99	10.00	10.18	10.38	10.95	11.16
11.95	11.99	12.75	13.18	13.49	14.15	17.21

(a) Construct a stemplot for these data, with levels 1 to 17. Identify three outliers on the stemplot.

(b) Construct a second stemplot for these data, with the outliers listed separately.

(c) Calculate the median from the stemplot in part (b).

(d) Obtain the extreme values and the range from the stemplot in part (b).

(e) Briefly comment on the distribution of this batch of data.

5 The shape of a batch

The stemplot is a very good method of picturing a batch of data because it enables you to see each individual data value (though these may be truncated) and, at the same time, to see the batch as a whole. It also shows the data values in numerical order, so you can easily find the 'middle' value (the median) and also the largest and smallest values (the upper and lower extremes). Perhaps its greatest strength, however, is that it shows you whether the values are spread out evenly or clustered together, and where this clustering takes place if so.

So a stemplot contains, in a readily accessible form, a lot of information about the *shape* of the values in a batch. Later in this section we shall look at ways of summarising such information. First, however, we shall look in more detail at how to prepare a useful stemplot of a batch of data, because this is frequently not as straightforward as it may have seemed so far.

5.1 How many levels?

In the stemplots you have drawn so far we have always told you the levels to put on the stem. Getting the number of levels right is important, since otherwise you may not get a useful impression of the overall shape of the batch. This is illustrated by the next example.

Example 14 Stemplots that are spread out or squashed up

The mileage readings, rounded to the nearest hundred miles, are collected for 22 cars on sale at a used car lot. The lowest mileage is 23 300 miles and the highest is 45 000 miles. If the stemplot is drawn with levels 23 to 45, so that each level represents 1000 miles, this gives the stemplot in Figure 21.

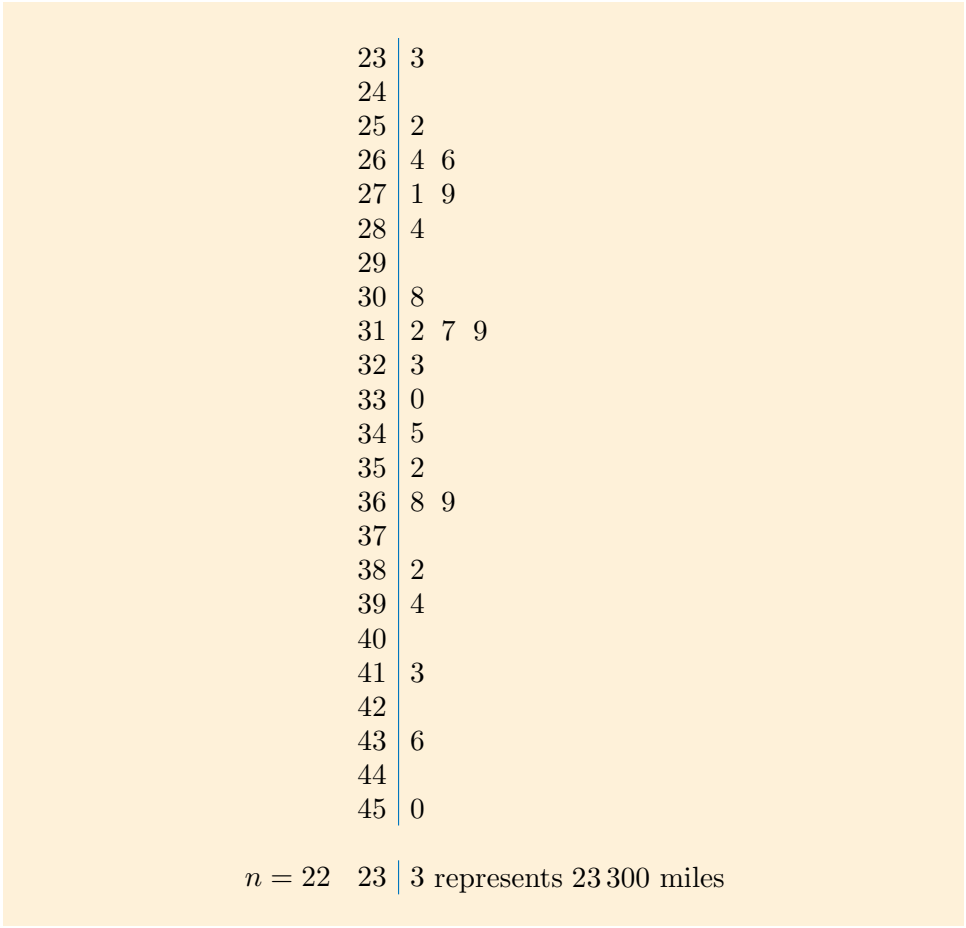


Figure 21 Stemplot of mileage of used cars (too many levels)

The values on the stemplot in Figure 21 are very spread out, which makes it difficult to get a good impression of the shape of the batch. The problem is that there are too many levels – there are 23 levels for 22 data values! So, the number of levels needs to be reduced in some way. Reducing them so that the levels are separated by 10 000 rather than 1000 miles produces the stemplot in Figure 22.

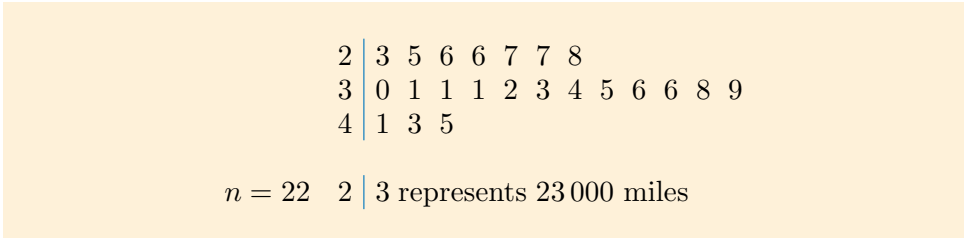


Figure 22 Stemplot of mileage of used cars (too few levels)

Now the stemplot is far too squashed up, so that, again, nothing much can be said about the shape of the batch. Reducing the number of levels from 23 to three is too drastic.

Choosing a suitable number of levels in a stemplot is a matter of trial and error. In this unit, we will always either tell you what levels to put on the stem of a stemplot, or let Minitab (a statistical package) do the work. However, there is one particular technique that you need to know: how to stretch a stemplot.

Minitab will be introduced in Section 6.

Example 15 Stretched stemplots

We want a way of stretching our squashed-up stemplot in Figure 22, by increasing the number of levels (without having too many). The idea of the **stretched stemplot** is to split each level into two or more parts. Thus, for instance, level 2, which currently can sprout leaves 0–9, is split into two parts, one bearing leaves 0–4 and one bearing leaves 5–9. Therefore, our stemplot with levels 2, 3 and 4 could now be stretched so as to have the following structure.

2	leaves 0–4
2	leaves 5–9
3	leaves 0–4
3	leaves 5–9
4	leaves 0–4
4	leaves 5–9

For example, level 3 of the stemplot in Figure 22 splits into two parts.

3	0 1 1 1 2 3 4
3	5 6 6 8 9

Splitting all the levels in this way gives the stretched stemplot in Figure 23, which now has six levels (or parts of levels).

2	3
2	5 6 6 7 7 8
3	0 1 1 1 2 3 4
3	5 6 6 8 9
4	1 3
4	5
$n = 22$ 2	3 represents 23 000 miles

Figure 23 Stretched stemplot of mileages of used cars

This is certainly an improvement over Figure 22, although it is rather squashed still, with little detail in the middle. Can we do better? Well, we can try – by stretching the stemplot a bit more and splitting each level into five rather than two parts, each with just two leaves. The stemplot will now have the following structure, with up to 15 parts.

2	leaves 0–1
2	leaves 2–3
2	leaves 4–5
2	leaves 6–7
2	leaves 8–9
3	leaves 0–1
3	leaves 2–3
3	leaves 4–5
3	leaves 6–7
3	leaves 8–9
4	leaves 0–1
4	leaves 2–3
4	leaves 4–5
4	leaves 6–7
4	leaves 8–9

For example, level 3 of the stemplot in Figure 22 now splits into five parts.

3	0 1 1 1
3	2 3
3	4 5
3	6 6
3	8 9

Splitting all the levels in this way gives the stretched stemplot in Figure 24, which has 12 levels (or parts of levels).

2	3
2	5
2	6 6 7 7
2	8
3	0 1 1 1
3	2 3
3	4 5
3	6 6
3	8 9
4	1
4	3
4	5

$n = 22$ 2 | 3 represents 23 000 miles

Figure 24 Another stretched stemplot of mileages of used cars

This gives a further, possibly better, representation of the shape of the batch. The values are quite evenly spread between 25 000 and 40 000 miles, with some clustering between 25 000 and 31 000 miles.

Example 15 is the subject of Screencast 4 for Unit 1 (see the M140 website).

Stretching stemplots gives you greater flexibility to choose the right number of levels for your data. The following activity will give you some practice at doing this.

Activity 17 Price of digital televisions

The following are the prices of 26 digital televisions with 22- to 26-inch LED screens, quoted online by a large department store in February 2012. The prices have been rounded to the nearest pound, to eliminate the distraction of having to deal with many prices ending in 9.99.

Table 11 Prices of digital televisions ($)

170	180	190	200	220	229	230	230	230
230	250	269	269	270	279	299	300	300
315	320	349	350	400	429	649	699	

(a) Construct a stemplot with levels 1, 2, 3 and 4, with the two high outliers listed separately.

(b) Stretch the stemplot you obtained in part (a) so that each level is split into two parts.

(c) Now stretch the stemplot you obtained in part (a) so that each level is split into five parts.

(d) Briefly comment on the different stemplots you have obtained.

Whether or not a squeezed or stretched stemplot is an improvement on the original stemplot is largely subjective. The aim is to obtain a stemplot which is easy to read, contains all the important information about the batch of data and reveals interesting or important patterns within it.

The next activity will give you some practice at obtaining a useful stemplot.

Activity 18 Long-jump championship results

Table 12 gives the results for 26 senior male athletes in a long-jump competition in the UK in June 2011. A long jump, pictured in Figure 25, is measured in metres.

Table 12 Long-jump championship results (metres)

6.53	6.44	7.38	4.36	6.99	4.68	6.96	5.60	6.72	6.24	7.15	6.41	6.64
6.81	6.05	6.73	5.92	6.45	6.37	6.94	6.26	6.59	6.35	6.52	6.36	6.93

(Data source: UK Athletics)

(a) Construct a stemplot of the data with levels 4 to 7.

(b) Stretch the stemplot you obtained in part (a) so that each level is split into two parts.

(c) Stretch the stemplot you obtained in part (a) so that each level is split into five parts, with two low outliers listed separately.

(d) Comment on the three stemplots you have obtained. Which stemplot do you think provides the most useful information?

Figure 25 Long jump athlete getting ready for landing

5.2 Peaks and symmetry

Figure 26 Mount Fuji, Japan

Once we have pictured a batch of data by preparing a stemplot, we can describe its **shape**. There are several things to look for, including the number of *peaks* and if there is any *symmetry* in the shape. (Figure 26 shows a symmetric 'peak'.)

Peaks

Batches of data are categorised according to the number of *clear* peaks they have. That is, levels (or parts of levels) that have more leaves than nearby levels (or parts of levels).

Example 16 Identifying the peaks

Figure 27 shows a stemplot of the moth counts discussed in Activity 12.

1	6 9
2	0 1 1 1 2 4 4 9
3	4 4 7 7 9
4	1 6 7 7 9
5	4
6	5 7
7	1

$n = 24$ 1 | 6 represents a count of 16

Figure 27 Stemplot of moth counts

The moth counts in this stemplot have one clear peak around level 2. Although level 6 has two leaves while levels 5 and 7 have only one leaf, this does not qualify level 6 as a clear peak. This is because even one extra leaf on level 5 would remove this peak. It's important not to over-interpret every small bump in the data as a peak, but to focus on the main features.

Sometimes the number of peaks is not so clear. For example, the wooden toy prices of Exercise 7 represented in the stemplot of Figure 28 (below) certainly cluster in the middle, around levels 7 to 10.

1	2 2
2	
3	
4	0
5	9
6	0
7	2 5 9 9
8	4 7
9	1 4 4 5 9
10	0 1 3 9
11	1 9 9
12	7
13	1 4
14	1
15	
16	
17	2

$n = 28$ 1 | 2 represents £1.20

Figure 28 Stemplot of wooden toy prices

There is a major peak at level 9 and perhaps a secondary peak at level 7, but no other clear peaks. The fact that levels 1 and 13 have two leaves is not worth mentioning, as small bumps in the data can occur by chance.

The statistical word for a peak is a **mode**. A batch with just one mode is called **unimodal**. Thus the first batch in Example 16 is clearly unimodal because it has just one peak. As mentioned above, we only count 'clear' peaks, and not small bumps which might be due to chance occurrences: it's important not to attach undue importance to such minor irregularities. However, as often is the case in statistics, what constitutes a 'clear' peak can be rather subjective!

Some data have more than one clear peak, as illustrated in the next example.

Example 17 Wind turbines

Data were collected on the power produced by a 100 kilowatt wind turbine, like that shown in Figure 29, during 44 periods of one hour each. It is usual to measure the output over an hour to smooth out fluctuations due to short severe gusts of wind. The stemplot of this batch of data is shown in Figure 30.

Figure 29 Wind turbines

0	0 0 0 0 0 0 0 0 0 0 0 0 0 2 2 5 6 6 7 8
1	5 6 6
2	2
3	3
4	7
5	6
6	5 8
7	0 1
8	1 3
9	2 8
10	1 2 2 3 4 4 5 5 6

$n = 44$ 10 | 6 represents 106 kilowatts

Usually we have used the first number in the stemplot to explain the meaning of the stem and leaves. In Figure 30 we have used the last, $10 \mid 6$, as $0 \mid 0$ is not very helpful.

Figure 30 Stemplot of wind-turbine power output

You can see that there are a lot of readings for level 0 from when there was little or no wind, and so there is a mode at 0. The turbine is designed not to exceed its maximum output, so the distribution is cut off sharply. However, there are many occasions when the turbine is close to producing its maximum power output, and so there is a second mode at about 100 kilowatts. The batch has two modes as the data values are concentrated at each end.

Example 17 presented data with two clear peaks. Such data are said to be **bimodal**. Data with three or more clear peaks are said to be **multimodal**. In the next activity you will have a go at identifying the number of peaks in a dataset.

Example 17 is the subject of Screencast 5 for Unit 1 (see the M140 website).

Activity 19 Hunting the bumps

How many modes are there in the following two datasets and how would you describe these data: unimodal, bimodal, or something else?

(a) Table 13 gives the coal production in the UK in 1970/71 in the 18 regions with coal-mines. Figure 31 is a stemplot of the regional coal production from Table 13.

Table 13 Coal production (thousand tonnes) by region, in UK coalmines 1970/71

Scottish North	5 283	North Western	6 132
Scottish South	5 892	North Derbyshire	9 777
Northumberland	6 272	North Nottingham	12 070
North Durham	5 111	South Nottingham	10 555
South Durham	7 551	South Midlands	8 859
North Yorkshire	9 481	Staffordshire	8 302
Doncaster	8 010	East Wales	6 899
Barnsley	7 794	West Wales	4 787
South Yorkshire	9 471	Kent	1 008

1	0
2	
3	
4	7
5	1 2 8
6	1 2 8
7	5 7
8	0 3 8
9	4 4 7
10	5
11	
12	0

$n = 18$ 1 | 0 represents 1 000 000 tonnes

Figure 31 Stemplot of UK coal production, 1970/71

(b) Figure 32 is a stemplot of the times taken in the finals of 400-metre races in England in June 2011.

47	7 8
48	0 2 2 7 8 9
49	0 4 6 6
50	8
51	1
52	
53	9
54	
55	1 3
56	6 6
57	2 2 3 4 8
58	3 9 9
59	2 4
60	1 2

$n = 31$ 47 | 7 represents 47.7 seconds

Figure 32 Stemplot of times in 400-metre races

Symmetry

If a horizontal line can be drawn across the stemplot of a batch so that the shape on one side is almost the mirror image of the shape on the other side then the batch is said to be **symmetric**.

Example 18 Symmetric or not?

Figure 33 shows the stemplot of percentage scores for 21 students in an English test.

0	5
1	2
2	9
3	1 3 7
4	0 1 4 6
5	2 6 7 9
6	4 8 9
7	1 7
8	6
9	3

$n = 21$ 0 | 5 represents a score of 5%

Figure 33 Stemplot of percentage scores in an English test

The batch of data in Figure 33 is virtually symmetric about the line drawn between levels 4 and 5, which is as near as possible to the median, so about half the values lie each side of it. (The median of this dataset is 52%.) The symmetry of the stemplot is not quite perfect, as level 2 has only one leaf while level 7 has two leaves, but small differences between the two sides can occur by chance and so the batch should be regarded as symmetric.

This is the same as Figure 19.

On the other hand, consider the stemplot of median salaries for the 20 highest-paid occupations, in Figure 34.

4	4 4 5 5 5 5 7 7 8
5	1 2 3 6 8 9
6	0
7	4 8
8	2
9	
10	
11	4

$n = 20$ 4 | 4 represents £44 000

Figure 34 Stemplot of median salaries for highest-paid occupations

This has a single peak at level 4, dropping down at higher salaries. The batch is clearly not symmetric.

A batch of data which is not symmetric is called **skew**.

A batch like that in Figure 34, where the large values are spread out and the small values are close together, is called **right-skew**. This name originates from another method of plotting data values, where the stemplot is effectively turned anticlockwise through a right angle, thus putting it on its side so that the small data values occur on the left and the large values on the right, as in Figure 35.

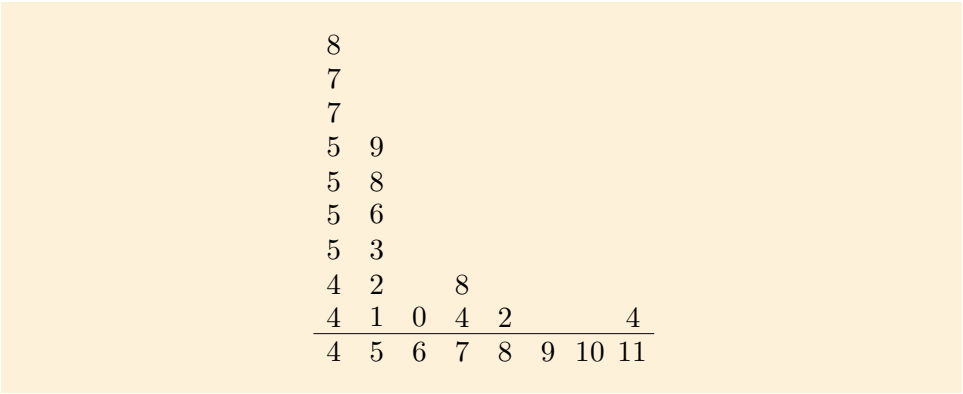


Figure 35 The stemplot of Figure 34 turned on its side

An equivalent way of thinking about skewness involves the **tails of a batch**. A batch has two tails: the lower tail is the set of values up to and including the largest mode (that is, the highest peak), and the upper tail is the set of values from the largest mode up. In a symmetric batch, the two tails are of similar length, and in a skew batch, one tail is more spread out than the other. If the batch is right-skew, the spread-out tail lies numerically above the mode; in Figure 35, it is to the *right*.

Another possibility in a non-symmetric batch is that the large values are close together and the low values are spread out. Such a batch is called **left-skew** because the spread-out tail would be on the *left* if the stemplot were turned on its side as described above.

Right skewness is also often referred to as *positive skewness* and left skewness as *negative skewness*.

Activity 20 Spotting the skew

Decide whether the batches in the following stemplots are symmetric, right-skew or left-skew.

(a) The moth-trap data from Activity 12.

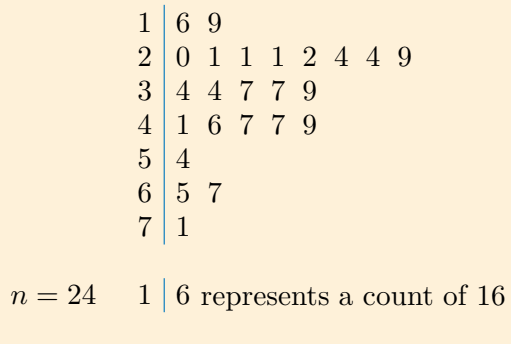


Figure 36 Stemplot of moth counts

(b) Percentage scores in arithmetic obtained by 33 school students.

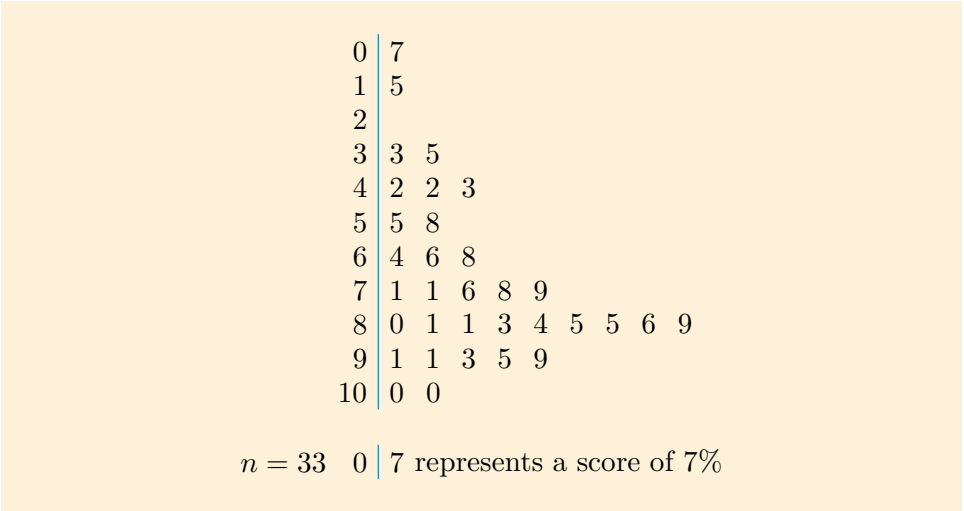


Figure 37 Stemplot of percentage scores in arithmetic

In this section, we have looked at ways of displaying and summarising a batch of data, so as to describe its shape in a useful manner. You have seen that it is important to choose the number of levels in a stemplot so that it is neither too spread out nor too cramped, and you have learned how to stretch a stemplot. You have also learned how to identify the modes of a batch of data and to say whether a batch is symmetric, right-skew or left-skew.

Exercises on Section 5

Exercise 8 Women's triple jump

Table 14 shows the performances of 16 senior female athletes in the triple jump during a UK championship in June 2011.

Table 14 Triple jump results (metres)

13.45	12.38	12.30	12.23	11.54	11.37	11.33	10.01
13.10	12.20	10.83	10.29	9.92	9.80	8.21	8.15

(Data source: UK Athletics)

(a) Construct a stemplot of these data with levels 8 to 13.

(b) Stretch the stemplot you obtained in part (a), by splitting each level into two parts.

(c) Comment briefly on the two stemplots. Identify two features of the data that are more apparent in the stretched stemplot.

Exercise 9 Describing shapes

For each of the two stemplots listed below, say how many modes there are, and describe the batch in terms of its symmetry or skew.

(a) The stemplot in Figure 38, showing the median salaries in the 20 lowest-paid occupations (discussed in Activity 15).

	2	1
	3	6
	4	
	5	5 6
	6	0 5 6 6 9
	7	0 8
	8	3 3 4 7 8
	9	1 6 9 9
$n = 20$	2	1 represents £2100

Figure 38 Stemplot of median salaries for the 20 lowest-paid occupations

(b) The stemplot in Figure 39, displaying the results of the shot-put competition described in Exercise 6 at the end of Section 4.

	11	9
	12	3 5 7
	13	8 9
	14	3
	15	1 2 4
	16	2 4 4 7
	17	
	18	0
$n = 15$	11	9 represents 11.9 metres

Figure 39 Stemplot of the shot-put data

6 Computer work: introducing Minitab

In this section, the statistical package Minitab is introduced via the Computer Book. You will learn how to use Minitab, including how to:

- perform calculations on data
- obtain stemplots
- print output
- paste output into another document.

You should now turn to the Computer Book and work through the Computer Book (html version)Introduction and Chapter 1.

7 Completing the assignments

There are two kinds of assignments in M140: interactive computer-marked assignments (iCMAs) and tutor-marked assignments (TMAs). The assignments and the instructions for submitting them are on the M140 website. This section provides some advice on how to complete these assignments, starting with some general advice on when to start working on them.

Most students find it best to start working on each assignment fairly soon after studying the material on which it is based. This usually means tackling the

assignment soon after studying the unit. It is usually not a good idea to defer starting work on an assignment until close to the cut-off date. This is because you may need some time to revise some topics, or contact your tutor with questions, and you are unlikely to produce your best work if you are under time pressure. Also, something unexpected might come up near the cut-off date, so you should allow some contingency time.

7.1 Answering iCMA questions

When you do the iCMA questions on a unit, you should have the unit and any notes that you made to hand. You will need a pen or pencil, paper, and your calculator. You will also need to go to the M140 website to access the iCMAs.

Make sure that you read each question of an iCMA carefully, so that you understand what is required before you start to work out your answer. You do not have to complete all the questions in an iCMA in one session; you can answer a few questions at a time, in any order, and save your answers. You can also change your answers in later sessions if you wish, before submitting the iCMA. Once you have completed the questions in an iCMA, it is a good idea to read through the questions again, to check that you are happy with your answers and that you have answered as many questions as you can. **Do not submit your iCMA until you are sure you have completed your work on it.** Once you have submitted your iCMA, you will not be able to retrieve it.

In the next activity, you are asked to try the practice quiz for this unit. The quiz is similar in style to the iCMA and the aim of the activity is to familiarise you with the process of answering iCMA questions, before attempting the first iCMA.

Activity 21 Doing the practice quiz for Unit 1

There are several versions of the practice quiz: the numbers in each question vary, but the questions are similar. When you start the practice quiz, you get a set of questions chosen at random.

Now go to the M140 website and access the practice quiz for Unit 1. Work through the questions and see how you get on. Unlike an iCMA, you can have as many attempts as you want with the practice quiz. So, if you get the wrong answer first time, you can always have another go – preferably after you've worked out what you did wrong! Return to this unit when you are done.

How did you get on? If you found the quiz difficult, then you may need to revise some topics in the unit, or contact your tutor for help. If you are happy with what you did, you can now start the first iCMA.

Activity 22 The first iCMA

Go to the M140 website and find the first iCMA. The first iCMA covers material from Units 1, 2 and 3, so **do not attempt the questions for Units 2 and 3 now**. Your answers are saved automatically when you complete a question, and you can log off and return to the iCMA as you wish. You can change your earlier answers up until the point you submit your iCMA.

You must submit your iCMA before the cut-off date or it will not count, but you can only submit the iCMA once. Hence you should only submit your iCMA once you have finished as much as you can, or if the cut-off date is upon you.

Follow the instructions given on the M140 website and try some of the iCMA questions relating to Unit 1. Complete as many of the questions for Unit 1 as you can. Return to this unit when you are done.

7.2 Answering TMA questions

TMAs are longer pieces of work than iCMAs. Unlike iCMAs, they enable your tutor to assess how you present and explain your statistical ideas, as well as the accuracy of your calculations. There are a few important points to remember when answering a TMA question, illustrated in the following examples and activities.

Example 19 shows why it is important that you show your workings, unless you are specifically asked not to. Good communication is an essential skill and you may lose marks if you do not show your workings.

Example 19 Showing your workings

Consider the following extract from a TMA question.

> The following table gives the lengths in centimetres (from tip of bill to tip of tail) for adult British birds of 30 different species.
>
9	9	12	12	12	13	14	14	14	15	15	16	16	18	21
23	23	24	25	26	27	32	32	34	41	46	46	47	47	64
>
> (a) Find the median length of these birds. [*2*]
>
> (b) Find the range. [*2*]

This question is fairly typical: there are some data, then you are asked to do two things with the data (calculate the median and the range). To the right, the numbers in square brackets tell you how many marks are allocated to each question. Generally, the more marks that are allocated, the more substantial your solution should be.

Consider part (a) of the question. The two marks allocated may include one mark for getting the method right, and one mark for doing the calculation correctly. A correct answer to this question might go something like:

> 'There are 30 data values, so the median is the average of the 15th and 16th largest values. These are 21 and 23. The average of these values is
>
> $$\frac{21+23}{2} = 22.$$
>
> Thus, the median is 22.'

This would get full (that is, two) marks: one for explaining the method and one for getting the right answer. Now consider the following answer:

> 'The median is 22.'

The answer is certainly correct, so the marker would presume that the right method has been used and therefore award the answer the full two marks. Note that if the question had also stated to 'show your workings', such an answer would *not* get full marks.

However, now consider the following answer:

'The median is 23.'

This answer is incorrect, so loses the accuracy mark. Furthermore, no workings have been shown, so it's impossible to tell whether the student has used the correct method and then made some arithmetic error. Thus such an answer would be awarded zero marks.

Finally, suppose that the student had shown their workings, and produced the following answer:

'There are 30 data values, so the median is the average of the 15th and 16th largest values. These are 21 and 25. The average of these values is

$$\frac{21+25}{2} = 23.$$

Thus the median is 23.'

The marker would see that the student has used the right method, but got the 16th largest value wrong. Therefore, the marker would award one mark for method and zero for accuracy. So the student would get one mark for this question.

Example 19 illustrates why it's a good idea to show your workings. In more complicated calculations, a small slip might not cost you all the accuracy marks – provided the marker can see where you've got it wrong. A further reason for showing your workings is that your tutor can then comment on where you went wrong, and provide you with useful feedback. This is much less likely to occur if you don't provide your tutor with something to go on.

Also, sometimes you may be given the answer, and asked to show that this is correct. In this case, all the marks are for method and explanation, and no marks are awarded for getting the correct answer, because it's *already* been given to you.

In the next activity, the tables are turned, and you are asked to mark a student's answer.

Activity 23 In your tutor's seat

Suppose you are a tutor marking a student's answer to part (b) of the TMA question in Example 19. You are required to award one mark for using the correct method, and one mark for numerical accuracy.

(a) You are told the correct answer is 55. Work this out for yourself.

(b) The student's answer is as follows:

'The range is the difference between the largest and the smallest values. These are 9 and 47, so the range is $47 - 9 = 38$.'

Mark the student's answer.

You will sometimes be asked to interpret your results. Typically, only brief answers are required. Occasionally, different interpretations may be possible, or different features may be picked out as noteworthy. The key point here is to support your interpretation with a brief explanation. This is shown in Example 20.

Example 20 Answering questions on interpretation

The following is a continuation of the TMA question described in Example 19.

The bird-size data are shown in the following stemplot.

0	9 9
1	2 2 2 3 4 4 4 5 5 6 6 8
2	1 3 3 4 5 6 7
3	2 2 4
4	1 6 6 7 7
5	
6	4

$n = 30$ 0 | 9 represents 9 cm

(c) How many modes are there in this batch? Give a reason for your answer. [2]

(d) Comment briefly on the shape of this batch. [2]

A suitable answer to part (c) might be as follows:

> 'There is a single mode at level 1. The small peak at level 4 is probably not significant. Thus, the batch is unimodal.'

This should get full marks, since the interpretation (the batch is unimodal) is supported by a reasonable argument (the peak at level 4 is insignificant). However, this *is* a matter of judgement. So here is another possible solution:

> 'There are two modes, at levels 1 and 4. While the peak at level 4 is much smaller than the one at level 1, it may nevertheless be significant. Thus, this batch is bimodal.'

This should also get full marks, because, again, the interpretation (the batch is bimodal) is supported by a reasonable argument (the peak at level 4 might be smaller than that at level 1, but it may still be significant). Statistics is like that: sometimes it can involve judgements that are subjective to some degree. However, the key thing is to support those judgements in such a way that the reader can assess their validity.

Note that you should write your answers out in full sentences, as shown in this example.

Part (d) is an open-ended invitation to comment, which should be kept brief. (It's only worth two marks.) An appropriate answer would be:

> 'The batch appears to be right-skew, since the higher values are more spread out than the lower values.'

This would get you both marks: one mark for correctly describing it as right-skew and one mark for saying why.

One thing to remember is that 'briefly' really does mean briefly!

Examples 19 and 20 are the subject of Screencast 6 for Unit 1 (see the M140 website).

Here is a final activity on these same data, to give you a bit more practice at answering open-ended questions.

Activity 24 An open-ended question

The stemplot in Example 20 can be improved by stretching it, so that each level is split into two.

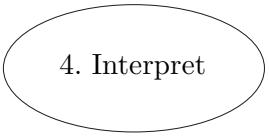
(a) Construct this stretched stemplot.

(b) Comment briefly on two key aspects of the stretched stemplot that you obtained in part (a).

A further important point in answering TMA questions is to write your answers out as full sentences, even when they include numerical or mathematical expressions. This is important for communicating your answers clearly. It is also a good idea to place each new step within a mathematical calculation on a new line.

One practical reason for communicating clearly is that, otherwise, your tutor might not be able to understand your reasoning, and thus you might lose marks that are awarded for using the correct method.

This advice is also important when reporting on analyses done using Minitab. It is not enough to enclose the Minitab output without further comment, and if you do you may lose marks.



A key step in the modelling diagram, introduced in Subsection 2.1, is to interpret the results of your analyses, and this involves conveying them in properly constructed sentences, not just a computer listing.

Example 21 A TMA question with Minitab

The following is an example of a TMA question that requires the use of Minitab.

(a) Enter the following data into a Minitab spreadsheet. [*1*]

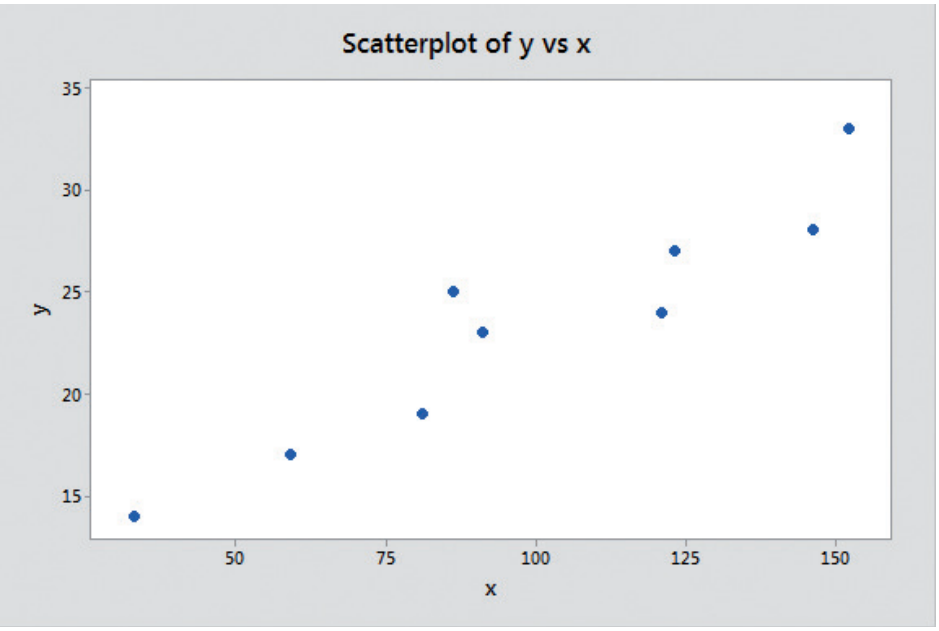
x	y
81	19
152	33
146	28
86	25
33	14
91	23
123	27
59	17
121	24

Use Minitab to produce a scatterplot of y against x. Include a copy of the scatterplot in the TMA answers that you send to your tutor. [*2*]

(b) Comment on the relationship between x and y. [*2*]

Part (a) requires you to use Minitab, and the evidence for what you have done will come from the scatterplot that you send to your tutor. You must make it clear which question the Minitab output relates to. (Do not send more output than is needed to answer the question.) Hence, a satisfactory answer to part (a) would be as follows.

'Scatterplot for part (a):



It is a scatterplot of y against x.'

Sending your tutor this scatterplot demonstrates that you know the method for entering data, which would gain one mark, and that you know the method for producing a scatterplot, which gains a second mark. The third mark would require the data to be accurate and the scatterplot to be drawn correctly (y against x and not x against y) with appropriate labels.

A reasonable answer for part (b) might say something like:

> 'A straight-line graph would be a good way of representing the relationship between y and x as the points lie roughly in a straight line.'

Note that your answer to (b) should be a complete sentence or complete sentences.

When you are working through the units and on your TMA, you are encouraged to seek clarification and help from your tutor or the module forums if you feel you need it. Also, you are encouraged to discuss your work on the module with other students, and to work together in groups if you find this helpful. However, your assignment answers should be your own work, and reflect your own thinking when you do them, even if that thinking has been informed by interactions with others. Your tutor will then be able to assess your progress and provide assistance tailored to your own personal needs if required. What you should not do, if you choose to work with others, is to submit 'group answers'. In addition, you should completely avoid copying answers, unacknowledged, from another source. **This would constitute plagiarism, which is treated severely at the Open University.**

Finally, remember that your tutor is there to offer guidance and support. For this reason, it is worth including partial solutions to questions that you haven't completed, and to send in your TMA even if it's incomplete. This could gain you some marks and some feedback, whereas if you send nothing in, you will receive neither of these.

The following box highlights some of the key points about doing TMAs.

Points to remember when answering TMA questions

- Show your workings – good communication is an important skill.
- Write your answers in full sentences.
- Place each new step in a calculation on a new line.
- Don't just paste computer output – make sure you interpret the output.
- When you interpret results, make sure you support your interpretation.
- Keep your answers brief.
- Your answers should be your own work, and reflect your own thinking.
- TMAs provide an opportunity for you to get marks and feedback, so it's worth sending in incomplete answers.

Summary

In this unit you have been introduced to the components of M140, and to some general ideas about statistical modelling. These are summarised in the modelling diagram, which involves four key steps: posing the question, collecting data, analysing the data, and interpreting the results. You have been reminded of these steps throughout the unit.

You have learned about the need to clean data before analysis, how to round data to specified numbers of decimal places and significant figures, and how to avoid spurious accuracy and rounding errors in calculations. Stemplots were introduced as an effective yet simple way of representing numerical data.

You learned how to construct stretched stemplots, how to recognise and present outliers, and also how to interpret the shape of a batch in terms of modes, symmetry and skewness. The median and range were described, and their resistance to outliers was discussed.

You learned how to use the statistical package Minitab, and how to do calculations on data and obtain stemplots with it. Finally, you worked through a practice quiz in preparation for doing the iCMA on the unit, and learned some key points about completing TMAs.

Learning outcomes

After working through this unit, you should be able to:

- follow the steps of a statistical investigation as set out in the modelling diagram
- recognise that data always need to be carefully inspected, and cleaned if necessary, before further analysis
- round numbers to a given accuracy
- round the final result of a calculation to avoid spurious accuracy
- draw a stemplot of a batch of data
- draw a stemplot in which outliers are listed separately
- draw a stretched stemplot, if appropriate
- use the stemplot of a batch of data to detect peaks (modes)
- use the stemplot of a batch of data to decide whether it is symmetric, left-skew or right-skew
- calculate the median of a batch of data using the stemplot
- find the upper and lower extremes of a batch
- calculate the range of a batch
- interpret the median as a summary measure of the location of a batch
- interpret the range as a summary measure of the spread of a batch
- use your computer to obtain, paste and print material from Minitab windows
- use Minitab to do numerical calculations on data
- use Minitab to obtain and customise stemplots.

Solutions to activities

Solution to Activity 1

No. There are many other possibilities. The next house could be in a different road and so have a completely different number. It may have just a name and no number. The next site could be empty, or perhaps the number 13 was deemed to be unlucky and so was skipped, so that the next house is number 15. Or it could be numbered 11A.

Solution to Activity 2

(a) The curve seems to become less steep as time goes on.

(b) One reason for this is that it is getting harder and harder to find new species of large marine animals. Of course, another reason could be that people are spending less time looking for them – although that seems unlikely. In fact, you'd expect that our ability to discover new species has improved over time.

(c) It seems likely that in the future the curve will become flat. You would expect this to happen once all species have been discovered, and the curve should then remain flat.

Solution to Activity 3

(a) The curve represents the predicted number of large marine species, and so the value 222.94 represents the predicted maximum number of such species. This number really ought to be a whole number, so we 'round it' to the nearest whole number. Thus, the predicted number of large marine species is 223.

(b) In 1996, based on this model, six large marine species were still to be discovered.

(c) These predictions are based on a particular model. Change the model, and the predictions may well change – either up or down. So, for example, if the curve is flattened a little towards the top, then the maximum value would reduce, whereas if the curve were a little steeper, the maximum number would increase.

Solution to Activity 4

Using the relationship between the three quantities, the expenditure for row 4 is

$$\frac{45.13 \times 115.9}{100} = 52.305\,67.$$

So the answer is *$*52.305 67.

Solution to Activity 5

(a) Rounding to the nearest million means rounding at the position indicated by the vertical bar:

$$6\,985\,|\,119\,415,$$

and the digit to the right of the rounding position is 1. Hence we round down to 6 985 000 000.

(b) This corresponds to rounding *up* to the nearest ten million.

Solution to Activity 6

(a) The rest of the data in the 'Expenditure' column in Table 2 are given to two decimal places, so we must round the calculated values to two decimal places as well. The rounded value in pounds is 52.31.

(b) The equation in Activity 4 can be rewritten as

$$\text{Petrol price (pence per litre)} = \frac{\text{Expenditure (£)} \times 100}{\text{Petrol used (litres)}}.$$

Substituting the values from row 28 of Table 2, using a calculator we obtain the value of the petrol price to be

$$\frac{54.06 \times 100}{40.37} = 133.911\,3203.$$

The other entries in the table are reported to one decimal place, hence the value should be rounded to one decimal place also. Thus the rounded petrol price in pence per litre is 133.9.

(c) The equation in Activity 4 can also be rewritten as

$$\text{Petrol used (litres)} = \frac{\text{Expenditure (£)} \times 100}{\text{Petrol price (pence per litre)}}.$$

Substituting the values from row 23 gives, on a calculator,

$$\frac{47.70 \times 100}{129.9} = 36.720\,554\,27.$$

Amounts of petrol are given to two decimal places in the table, so two decimal places should be kept, giving the rounded answer in litres as 36.72. The calculation for row 32 gives

$$\frac{57.43 \times 100}{132.9} = 43.212\,942\,06,$$

so the rounded value in litres is 43.21 to two decimal places.

Solution to Activity 7

Multiplying the two numbers, and then rounding to a whole number, gives

$$19.4 \times 23.4 = 453.96 \simeq 454.$$

Rounding first (to 19 and 23) and then multiplying gives

$$19 \times 23 = 437.$$

There is a difference of 17 between these two results. This is a big difference and so this degree of rounding error is unacceptable. The correct result is 453.96, which rounded to a whole number gives 454.

Solution to Activity 8

The mileage was 112 350 on 18.01.10 and 112 616 on 18.02.10, so the distance travelled between these dates was $112\,616 - 112\,350 = 266$ miles.

Solution to Activity 9

(a) 2460 (rounded to the nearest ten) has three significant figures: 2, 4 and 6.

(b) 0.003 610 rounded to six decimal places has four significant figures: 3, 6, 1 and 0.

(c) If the number has been rounded to the nearest ten, then there are two significant figures: 9 and 1. If it has been rounded to the nearest whole number (for example, from 909.8), then there are three: 9, 1 and 0.

(d) To keep three significant figures we round at the positions indicated by the vertical bars:

$$208|.3 \quad \text{and} \quad 0.0983|765.$$

Therefore, the rounded numbers are 208 and 0.0984.

Solution to Activity 10

From Table 2 the mileage value on 18.02.10 was 112 616 and on 04.03.10 it was 112 954. The number of miles travelled between these dates is therefore $112\,954 - 112\,616 = 338$.

The amount of petrol bought on 04.03.10 was 43.49 litres. From the calculator, the quantity of petrol in gallons bought on that date is therefore $43.49/4.546\,09 = 9.566\,462\,609$.

This figure is also the volume of petrol used between the two dates (since the tank was filled up at each petrol stop). The petrol consumption in miles per gallon is therefore:

$$338/9.566\,462\,609 = 35.331\,764\,08,$$

on a calculator. The numbers involved in the calculation, 338, 43.49 and 4.546 09, have three, four and six significant figures respectively, and can all reasonably be considered as having been rounded. Hence we must round our result so that it has three significant figures, so the correctly rounded petrol consumption is 35.3 miles per gallon.

Solution to Activity 11

(a) The stemplot has 17 levels, numbered 26 to 42.

(b) This means that the batch had no values between 27.0 and 27.9, or between 41.0 and 41.9.

(c) Seven levels have a single leaf: levels 30, 31, 33, 34, 37, 40 and 42.

(d) The value 35.3 occurred twice in the batch: the leaf '3' occurs twice at level 35.

Solution to Activity 12

(a) There are no count values below 16 and no count values above 71. Levels 1 to 7 inclusive will therefore suffice. (Level 0 would be needed for counts less than 10, while levels 8 and above would be needed for counts of 80 and more.)

(b) The following figure is the stemplot that results when you proceed from left to right on row 1 of the data, putting the leaves on the stemplot as you go, and then doing the same thing for row 2. (No need to add the key at the bottom for now.)

1	6 9
2	1 4 1 0 1 9 4 2
3	9 4 4 7 7
4	7 9 6 1 7
5	4
6	5 7
7	1

Unordered stemplot for moth-count data

(c) Ordering the leaves on each level, so that they are in increasing numerical order as you move away from the stem, gives the following stemplot.

1	6 9
2	0 1 1 1 2 4 4 9
3	4 4 7 7 9
4	1 6 7 7 9
5	4
6	5 7
7	1

Ordered stemplot for moth-count data

(d) There are 24 data values, so the batch size is $n = 24$. The key is: $1 \,|\, 6$ represents a count of 16. So the following line should be added to the stemplot:

$n = 24 \quad 1 \,|\, 6$ represents a count of 16.

Solution to Activity 13

From the stemplot, you can see that there is a clear tendency for the data values to cluster around level 2, which corresponds to counts in the twenties, with fewer values at higher and lower levels.

The moth counts tail off more gradually towards the higher values, and drop more suddenly towards the lower values.

Solution to Activity 14

From the ordered stemplot the batch size is 24. As this is an even number we must find the middle *two* values, which are the 12th and 13th values. The 12th value, counting from the lowest value, is 34 and the 13th value is 37. So the median value is $(34 + 37)/2 = 35.5$. We must round this to the same level of accuracy as the data values, so the median value is 36.

Solution to Activity 15

(a) Since the levels are 2 to 9, and one leaf digit is retained, we drop the last two digits of each number (without rounding). The data used to construct the stemplot are in the table below. The corresponding stemplot follows after. The outliers are marked *LO* and are 21, corresponding to *$*2100, and 36, corresponding to *$*3600.

21	36	55	56	60	65	66	66	69	70
78	83	83	84	87	88	91	96	99	99

LO		21 36
5	5	6
6	0	5 6 6 9
7	0	8
8	3	3 4 7 8
9	1	6 9 9
$n = 20$	5	5 represents £5500

Stemplot of median salaries for lowest-paid occupations, with low outliers

(b) The batch size is 20, so the median is the average of the 10th and 11th largest values. These are *$*7000 and *$*7800, so the average is *$*7400. (Based on the original data values, the median is *$*7450.)

(c) If the two lowest groups saw their median salaries rise by *$*1000, this would not affect the ordering. Therefore, the 10th and 11th values would not be affected and hence the median would stay the same.

Solution to Activity 16

For this batch, $E_U = 114\,000$ and $E_L = 44\,000$ (from the stemplot in Figure 19). Therefore,

$$E_U - E_L = 114\,000 - 44\,000 = 70\,000.$$

So the range is *$*70 000.

Solution to Activity 17

(a) Dropping the final digit gives the following data. There are two outliers, 64 and 69. Listing these separately, the stemplot has levels 1–4, as shown below.

17	18	19	20	22	22	23	23	23
23	25	26	26	27	27	29	30	30
31	32	34	35	40	42	64	69	

1	7 8 9
2	0 2 2 3 3 3 3 5 6 6 7 7 9
3	0 0 1 2 4 5
4	0 2
HI	64 69

$n = 26$ 1 | 7 represents £170

Stemplot of television prices

(b) Splitting each level into two parts gives the following stretched stemplot.

1	7 8 9
2	0 2 2 3 3 3 3
2	5 6 6 7 7 9
3	0 0 1 2 4
3	5
4	0 2
HI	64 69

$n = 26$ 1 | 7 represents £170

Stretched stemplot of television prices

(c) Splitting each level into five parts gives the following stretched stemplot.

1	7
1	8 9
2	0
2	2 2 3 3 3 3
2	5
2	6 6 7 7
2	9
3	0 0 1
3	2
3	4 5
3	
3	
4	0
4	2
HI	64 69

$n = 26$ 1 | 7 represents £170

Another stretched stemplot of television prices

(d) The stemplot in part (a) is perhaps a little too cramped, and the stemplot in part (c) a little too spread out. The stemplot in part (b) satisfies the 'Goldilocks' principle – that is, it's just right! The prices of these televisions cluster around *$*200–*$*300, with a couple of high outliers.

Solution to Activity 18

(a) The required stemplot is shown below.

4	3	6
5	6	9
6	0	2 2 3 3 3 4 4 4 5 5 5 6 7 7 8 9 9 9 9
7	1	3

$n = 26$ 4 | 3 represents 4.3 metres

Stemplot of long jumps

(b) Splitting each level into two parts yields the following stretched stemplot.

4	3	
4	6	
5		
5	6	9
6	0	2 2 3 3 3 4 4 4
6	5	5 5 6 7 7 8 9 9 9 9
7	1	3

$n = 26$ 4 | 3 represents 4.3 metres

Stretched stemplot of long jumps

(c) Splitting each level into five parts gives the stretched stemplot below.

4	3	
4		
4	6	
4		
5		
5		
5		
5	6	
5	9	
6	0	
6	2	2 3 3 3
6	4	4 4 5 5 5
6	6	7 7
6	8	9 9 9 9
7	1	
7	3	

$n = 26$ 4 | 3 represents 4.3 metres

Another stretched stemplot of long jumps

Then, by listing the two low outliers separately, we end up with the subsequent stemplot.

LO	43 46
5	6
5	9
6	0
6	2 2 3 3 3
6	4 4 4 5 5 5
6	6 7 7
6	8 9 9 9 9
7	1
7	3
$n = 26$ 5	6 represents 5.6 metres

Final stemplot of long jumps

(d) The final stemplot is probably the most useful. The stemplot from part (a) is far too cramped, with a large number of values at level 6. The stemplot from part (b) is much better, but still very crowded at the higher levels. The final stemplot gives much more detail.

Solution to Activity 19

(a) Hardly any areas produced less than 4000 thousand tonnes or more than 10 000 thousand tonnes. Most areas produced between 5000 thousand and 9000 thousand tonnes, and somewhere in this range it's possible that there is a single peak, though it's not very strongly marked. If this is correct, then these data are unimodal. If not, then the answer is 'something else'! There's a suggestion that the data are bimodal, though again the peaks are not clear. The honest answer is that in this case, it's rather hard to tell how many modes there are.

(b) These data have one peak at around 48 seconds and another at around 57 seconds, so they are bimodal.

In fact, these times were for both male and female athletes and you can clearly see the two distributions. When there is more than one mode, it is often the case that there are subgroups within the data that can explain the different peaks.

Solution to Activity 20

(a) In the moth-trap data, the larger values are more spread out than the smaller values, so the batch is right-skew.

(b) In the arithmetic test data, the smaller values are more spread out than the larger values, so the batch is left-skew.

Solution to Activity 23

(a) The range is the difference between the largest and the smallest values. These are 64 and 9, so the range is $64 - 9 = 55$.

(b) The numerical value of the answer is wrong, so the student should lose the accuracy mark. However, the method is correct, and the student has just got the wrong 'largest' value. So this answer should be awarded one mark out of the two available.

Solution to Activity 24

(a) The stretched stemplot is as follows.

	0	9 9
	1	2 2 2 3 4 4 4
	1	5 5 6 6 8
	2	1 3 3 4
	2	5 6 7
	3	2 2 4
	3	
	4	1
	4	6 6 7 7
	5	
	5	
	6	4
$n = 30$	0	9 represents 9 cm

(b) An appropriate answer would be:

> 'The stemplot is bimodal, with a peak at level 1 and a smaller peak at level 4. The distribution is not symmetric, but it appears to be right-skew, since the higher values are more spread out than the lower values.'

Note that stretching out the stemplot makes it clearer that there may indeed be a second mode – the presence of a second mode is much less apparent in the unstretched stemplot. In an open-ended question like this, there is some scope for what points to focus on. For example, you might choose to mention the possible presence of a high outlier at 64. This would count as another valid 'key aspect'.

Solutions to exercises

Solution to Exercise 1

Your answers may look something like this:

- Collect data: go through past bank statements and extract all items of household expenditure over the past few months, indicating the date, the item, and the amount.
- Analyse data: arrange expenditure items under convenient headings (bills, food, travel, entertainment, etc.), sum them up and then obtain monthly average expenditure totals under each heading.
- Interpret results: critically examine your past expenditure under each heading, and decide upon a strategy to make savings in the future.

Solution to Exercise 2

(a) The discovery curve becoming gradually less steep reflects the fact that times between successive discoveries is increasing as the numbers of species remaining to be discovered decreases.

(b) This could perhaps be used to predict how long it might be expected to take before the next species is discovered. The accuracy of such a prediction might not be very good as improvements in detection techniques might reduce the time until the next discovery. Or, these last few species might be hiding away! And of course, if the last large marine species discovered is in fact the last existing one, the time until the discovery of the next will be infinite.

Solution to Exercise 3

(a) 502.562

(b) 502.561 53

(c) 503

(d) 500.

Solution to Exercise 4

(a) 6 985 120 000, which has six significant figures

(b) 6 985 100 000, which has five significant figures.

Solution to Exercise 5

The petrol used between those two dates is 46.01 litres. The distance travelled is $112\,616 - 112\,350 = 266$ miles. Thus the consumption in litres per 100 kilometres is

$$\frac{100 \times 46.01}{266 \times 1.609\,344} = 10.747\,852\,84.$$

Now 46.01 has four significant figures, 266 has three and the conversion factor has seven. All have been rounded to some extent. The multiplier 100 (which has just one significant figure), on the other hand, has not been rounded, so does not count. So three significant figures should be kept in the final result, which is therefore 10.7 litres per hundred kilometres.

Solution to Exercise 6

(a) Dropping the last digit of each value (without rounding), we obtain the following stemplot.

11	9
12	3 5 7
13	8 9
14	3
15	1 2 4
16	2 4 4 7
17	
18	0

$n = 15$ 11 | 9 represents 11.9 metres

Stemplot for the shot-put data

(b) There are 15 data points, so the median is the 8th largest value. This is 15.1 metres.

(c) The extreme values are $E_L = 11.9$ and $E_U = 18.0$. Therefore, the range is $18.0 - 11.9 = 6.1$ metres.

(d) The distribution clusters around two sets of levels: levels 12–13, and levels 15–16. The maximum value is separated from the rest by a level with no leaves, so could perhaps qualify as an outlier.

Solution to Exercise 7

(a) The following is the first stemplot. There are three outliers: two low outliers at level 1, both corresponding to *$*1.20, and one high outlier at level 17 corresponding to *$*17.20.

1	2 2
2	
3	
4	0
5	9
6	0
7	2 5 9 9
8	4 7
9	1 4 4 5 9
10	0 1 3 9
11	1 9 9
12	7
13	1 4
14	1
15	
16	
17	2

$n = 28$ 1 | 2 represents £1.20

Stemplot of wooden toy prices

(b) The following is the stemplot with the outliers listed separately.

LO	12 12
4	0
5	9
6	0
7	2 5 9 9
8	4 7
9	1 4 4 5 9
10	0 1 3 9
11	1 9 9
12	7
13	1 4
14	1
HI	172

$n = 28$ 1 | 2 represents £1.20

Stemplot of wooden toy prices, with outliers listed separately

(c) There are 28 data values, so the median is the average of the 14th value, 9.4, and the 15th value, 9.5. The average of these two values is 9.45. Rounding to the same accuracy as the other values on the stemplot gives the median to be 9.5.

(d) The lower extreme is 1.2, and the upper extreme is 17.2, so the range is 16.0.

(e) While prices of toys under *$*20 vary widely, most cluster in the region *$*7–*$*11.

Solution to Exercise 8

(a) We drop the last digit (without rounding) and obtain the following stemplot.

8	1 2
9	8 9
10	0 2 8
11	3 3 5
12	2 2 3 3
13	1 4

$n = 16$ 8 | 1 represents 8.1 metres

Stemplot of triple-jump results

(b) Splitting each level into two parts gives the following stretched stemplot.

8	1 2
8	
9	
9	8 9
10	0 2
10	8
11	3 3
11	5
12	2 2 3 3
12	
13	1 4

$n = 16$ 8 | 1 represents 8.1 metres

Stretched stemplot of triple-jump results

(c) The stretched stemplot in part (b), unsurprisingly, is more spread out. It reveals a clustering of values between 9.5 and 12.4 metres which is not immediately apparent in the standard stemplot from part (a). Also, the stretched stemplot reveals some potential low and high outliers, which were not apparent on the standard stemplot.

Solution to Exercise 9

(a) There appear to be two modes, one at level 6 and one at level 8. Thus, these data are bimodal. The data are not symmetric, since the lower values are more spread out than the higher values. The data are left-skew.

(b) There are two modes, one at level 12 and one at level 16. Thus these data are bimodal. If we ignore the high outlier at level 18, the data appear roughly symmetrical around level 14.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Cover image: Minxlj/www.flickr.com/photos/minxlj/422472167/. This file is licensed under the Creative Commons Attribution-Non commercial-No Derivatives Licence http://creativecommons.org/licenses/by-nc-nd/3.0/

Figure 1 Taken from: http://en.wikipedia.org/wiki/File:Mort.svg. This file is licensed under the Creative Commons Attribution-Share Alike Licence http://creativecommons.org/licenses/by-sa/3.0/.

Figure 2 Royal Statistical Society

Figure 6 Richard Ling. http://en.wikipedia.org/wiki/File:Blue_Linckia_Starfish.JPG. This file is licensed under the Creative Commons Attribution-Share Alike Licence http://creativecommons.org/licenses/by-sa/3.0/.

Figure 10 Ilfremer / A.Fifis

Figure 12 With permission from: www.rac.co.uk

Figure 15 Shogo Kato

Figure 17 Trevor & Dilys Pendleton. www.eakringbirds.com

Figure 18 Mike Kemp/In Pictures/Corbis

Figure 25 Taken from: http://en.wikipedia.org/wiki/File:Jessica_Ennis_-_long_jump_-_3.jpg. This file is licensed under the Creative Commons Attribution-Noncommercial-ShareAlike Licence http://creativecommons.org/licenses/by-nc-sa/3.0/.

Figure 29 Taken from: http://thefutureofthings.com/upload/items_icons/Repower-5M-wind-turbine_large.jpg

Subsection 3.2 figure, ‘Rounding up’, Christopher Furlong / Getty Images

Subsection 5.2 figure, ‘Leaning Tower of Pisa’ © Jean-Yves Benedeyt.

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked the publishers will be pleased to make the necessary arrangements at the first opportunity.

Index